Docket No. 826.1540/JDH

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Hiroshi TSUDA

Serial No.: To Be Assigned

Filed: March 12, 1999

Group Art Unit: To Be Assigned

Examiner: To Be Assigned

For: DOCUMENT FILE GROUP ORGANIZING APPARATUS AND METHOD THEREOF

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. § 1.55

Assistant Commissioner for Patents
Washington, D.C. 20231

Sir:

In accordance with the provisions of 37 C.F.R. § 1.55, Applicants submit herewith a certified copy of each of the following foreign application:

Japanese Appln. No. 10-176749, filed June 24, 1998.

It is respectfully requested that Applicants be given the benefit of the earlier foreign filing date, as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. § 119.

Respectfully submitted,

STAAS & HALSEY

Dated: March 12, 1999

By: _____

James D. Halsey, Jr.
Registration No. 22,729

700 Eleventh Street, N.W.
Suite 500
Washington, D.C. 20001
(202) 434-1500

PATENT OFFICE

JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: June 24, 1998

Application Number: Patent Application No. 10-176749

Applicant(s): FUJITSU LIMITED

October 9, 1998

Commissioner,

Patent Office Takeshi Isayama

Certificate No. 10-3081505

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９８年　６月２４日

出願番号
Application Number: 平成１０年特許願第１７６７４９号

出願人
Applicant (s): 富士通株式会社

１９９８年１０月　９日

特許庁長官
Commissioner,
Patent Office 伊佐山建志

出証番号　出証特平１０－３０８１５０５

【書類名】　特許願

【整理番号】　9800955

【提出日】　平成10年 6月24日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 17/30

【発明の名称】　文書整理装置および方法

【請求項の数】　17

【発明者】

【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

【氏名】　津田　宏

【特許出願人】

【識別番号】　000005223

【氏名又は名称】　富士通株式会社

【代理人】

【識別番号】　100074099

【郵便番号】　102

【住所又は居所】　東京都千代田区二番町８番地２０　二番町ビル３Ｆ

【弁理士】

【氏名又は名称】　大菅　義之

【電話番号】　03-3238-0031

【選任した代理人】

【識別番号】　100067987

【郵便番号】　222

【住所又は居所】　神奈川県横浜市港北区太尾町１４１８－３０５（大倉山二番館）

【弁理士】

【氏名又は名称】　久木元　彰

【電話番号】　045-545-9280

【手数料の表示】

【予納台帳番号】　012542

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9705047

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書整理装置および方法

【特許請求の範囲】

【請求項１】　文書群をキーワードに基づいて整理する文書整理装置であって、

与えられたキーワードからキーワード間の階層関係と連想関係を抽出する関係抽出手段と、

前記階層関係と連想関係をリンクとして用いて、前記文書群にアクセスするためのディレクトリ情報を生成する生成手段と、

前記ディレクトリ情報を出力する出力手段と

を備えることを特徴とする文書整理装置。

【請求項２】　前記関係抽出手段は、前記与えられたキーワードからキーワード対を含む相関ルールを抽出するルール抽出手段と、該相関ルールを評価して、前記階層関係、前記連想関係、および同値関係のいずれかを該キーワード対に付与するルール評価手段を含むことを特徴とする請求項１記載の文書整理装置。

【請求項３】　前記ルール抽出手段は、共起出現頻度の高いキーワード対を前記相関ルールとして抽出し、前記ルール評価手段は、抽出されたキーワード対の関係を該共起出現頻度に基づいて付与することを特徴とする請求項２記載の文書整理装置。

【請求項４】　前記関係抽出手段は、前記与えられたキーワードから共起出現頻度の高いキーワード対を抽出する手段と、該キーワード対の共起出現頻度に基づいて、前記階層関係、前記連想関係、および同値関係のいずれかの関係を該キーワード対に付与する手段を含むことを特徴とする請求項１記載の文書整理装置。

【請求項５】　前記関係抽出手段は、前記与えられたキーワードから部分文字列を切り出し、該与えられたキーワードと部分文字列の包含関係を前記階層関係として抽出する切り出し手段を含むことを特徴とする請求項１記載の文書整理装置。

【請求項６】　前記関係抽出手段は、キーワード対の共起出現頻度に基づい

て第１の階層関係を抽出する手段と、前記与えられたキーワードと部分文字列の包含関係から第２の階層関係を抽出する手段と、該第１および第２の階層関係と外部から与えられた階層関係とをマージする手段とを含むことを特徴とする請求項１記載の文書整理装置。

【請求項７】　前記生成手段は、前記階層関係と連想関係を用いて、トップカテゴリからディレクトリへのパス、該ディレクトリの上位関連語、該ディレクトリのサブカテゴリ、該ディレクトリの連想語、および５０音・アルファベット順索引のうち、少なくとも１つ以上のリンクを含むハイパーテキストによる索引を前記ディレクトリ情報として生成し、前記出力手段は、前記文書群を該索引により整理して出力することを特徴とする請求項１記載の文書整理装置。

【請求項８】　前記生成手段は、前記階層関係と連想関係を用いて、トップカテゴリのキーワードから各キーワードへのパスを自動的に計算し、得られたパスを前記トップカテゴリからディレクトリへのパスとして設定することを特徴とする請求項７記載の文書整理装置。

【請求項９】　外部から与えられた同意語リストに基づいて、キーワード間の同値関係を付加する手段をさらに備え、前記生成手段は、該同値関係を含むディレクトリ情報を生成することを特徴とする請求項１記載の文書整理装置。

【請求項１０】　外部から与えられた不要語リストに基づいて、キーワードを削除する手段をさらに備え、前記生成手段は、削除されたキーワードを除くディレクトリ情報を生成することを特徴とする請求項１記載の文書整理装置。

【請求項１１】　外部から与えられたキーワード間の階層関係を入力する手段をさらに備え、前記生成手段は、該与えられたキーワード間の階層関係を用いてディレクトリ情報を生成することを特徴とする請求項１記載の文書整理装置。

【請求項１２】　前記与えられたキーワードを旧キーワードと比較して、新規キーワードを同定する手段をさらに備え、前記出力手段は、該新規キーワードを強調して出力することを特徴とする請求項１記載の文書整理装置。

【請求項１３】　前記ディレクトリ情報にアクセスするアクセス手段をさらに備え、利用者は、該ディレクトリ情報を介して前記文書群にアクセスすることを特徴とする請求項１記載の文書整理装置。

【請求項１４】　前記ディレクトリ情報に含まれるキーワードを検索するキーワード検索手段と、前記文書群の文書の内容を検索する文書検索手段をさらに備え、利用者は、該キーワード検索手段および文書検索手段を用いて文書情報を取得することを特徴とする請求項１記載の文書整理装置。

【請求項１５】　任意の情報をキーワードに基づいて整理する情報整理装置であって、

与えられたキーワードからキーワード間の階層関係と連想関係を抽出する関係抽出手段と、

前記階層関係と連想関係をリンクとして用いて、前記任意の情報にアクセスするためのディレクトリ情報を生成する生成手段と、

前記ディレクトリ情報を出力する出力手段と

を備えることを特徴とする情報整理装置。

【請求項１６】　文書群をキーワードに基づいて整理するコンピュータのためのプログラムを記録した記録媒体であって、

与えられたキーワードからキーワード間の階層関係と連想関係を抽出するステップと、

前記階層関係と連想関係をリンクとして用いて、前記文書群にアクセスするためのディレクトリ情報を生成するステップと

を含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項１７】　与えられたキーワードからキーワード間の階層関係と連想関係を抽出し、

前記階層関係と連想関係をリンクとして用いて、文書群にアクセスするためのディレクトリ情報を生成し、

前記ディレクトリ情報に基づいて前記文書群を整理する

ことを特徴とする文書整理方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、情報処理装置に蓄えられた大量の文書ファイル群を、その内容に基づいて整理する文書整理装置およびその方法に関する。

【０００２】

【従来の技術】

今日、コンピュータネットワークの発達により、大量のオンライン文書情報が溢れてきており、文書検索および文書整理に対する利用者の期待も大きくなっている。例えば、インターネットのホームページ検索サービスでは、大別して次の２種類のサービスまたはそれらの組合せが提供されている。

（ａ）ディレクトリサービス系

ホームページを階層的に分類して整理する。

（ｂ）全文検索系

ロボット（検索プログラム）が集めたページの全文検索を行う。

【０００３】

ある有名なディレクトリサービスでは、ディレクトリの作成にあたって、次のような方法を取っている。

１．ホームページ作成者が、ホームページを登録したいＵＲＬ（uniform resource locator）を申請する。

【０００４】

２．サービス提供者が、ホームページを階層カテゴリに分類して登録する。

３．階層カテゴリはサービス提供者独自のものであり、常に変化する。また、１つのホームページは複数のカテゴリに分類される。

【０００５】

このサービスでは、サーファと呼ばれる十数名の専門家がディレクトリの作成や情報のメンテナンスを行っており、これにより、常に良質な情報を提供することができる。しかし、大量の文書を分類するための人手をコンスタントに確保するのは、実際には難しい。また、個人が受信した大量の電子メール等を分類するような場合、手作業でディレクトリの作成等を行うのは非常に煩雑である。そこで、コンピュータによる文書自動分類の研究が期待されている。

【０００６】

従来、分類学（タクソノミ）では、情報が木構造により分類され、木の分岐において、各子ノードは互いに独立である。また、交叉分類は許されず、情報の配置位置は木構造上の１箇所に限られるという特徴がある。

【０００７】

このようなタクソノミの方法を文書検索に用いる場合、文書が木構造により分類され、１つの文書へのパスはただ一つに限られる。しかし、利用者が分類者と同じような分類基準を持っているとは限らないので、文書に辿りつくのが困難な場合があり、必ずしも有効な方法とは言えない。

【０００８】

そこで、分類を文書検索に利用する場合には、１つの文書に複数のカテゴリを与えることが考えられる。例えば、インターネットのディレクトリサービスのディレクトリ構造はそのようになっている。また、「文書情報分類方法および文書情報分類装置」（特開平８－１５３１２１）では、文書群のキーワードから階層的なカテゴリを作成し、文書を複数のカテゴリに格納している。

【０００９】

ところで、文書自動分類の研究では、大別して次の２つのアプローチがある。どちらも、利点や欠点があり用途によって使い分けたり、両者を組み合わせたりする必要がある。

（ａ）クラスタリング

キーワードの統計／表層上の関係をベースにして、与えられた文書群をいくつかの適当なクラスに分割する。このアプローチの利点は、既存の分類にとらわれずに、元の文書群の特徴を反映した分類結果が得られることであり、欠点は、自動化の精度が低いことである。

（ｂ）カテゴライゼーション

与えられた文書が既存の分類におけるどのカテゴリに当てはまるかを判別する。既存の分類としてはシソーラス（概念階層辞書）等があり、文書中のキーワードの分布等を手掛かりにして、その文書を近いカテゴリに当てはめていく。このアプローチの利点は、自動化の精度がクラスタリングより高いことであり、欠点は、分類結果が汎用的で、元の文書群の特徴を反映しないことである。

【００１０】

インターネットのディレクトリサービスの多くでは、通常、既存の分類に対してカテゴライゼーションを人手で行う。そして、１つのクラスが大きくなった場合には、クラスを分割するクラスタリング操作を、やはり人手で行っている。

【００１１】

例えば、上述の特開平８－１５３１２１のシステムは、文書に付加されたキーワードに基づいてクラスタリングを行うシステムである。また、キーワードによるクラスタリングの欠点を補うため、汎用のシソーラスを用い、意味属性の統計を利用して分類精度を上げる研究も報告されている（河合　敦夫，意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌，Vol.33，No.9，pp.1114-1122，1992.）。

【００１２】

【発明が解決しようとする課題】

しかしながら、上述した従来の文書分類システムには、次のような問題がある。

【００１３】

まず、人手による分類は、ディレクトリを作成したり、運用したりする専門家を必要とし、一般の利用者が分類を行うことは難しい。また、ディレクトリのハイパーテキストのメンテナンスを人手で行う場合、管理者の手間が大きく、単純なミスも発生しやすい。

【００１４】

また、タクソノミに従って文書を自動分類すると、通常、情報は１つのカテゴリにしか分類されない。この場合、利用者が分類者と同じような分類基準を持っていなければ、情報に辿りつくのが困難となることがある。また、クラスタリング、カテゴライゼーションともに完全な自動化は不可能であり、分類にゴミやもれがあると、利用者が情報に辿りつくのはさらに困難となる。

【００１５】

さらに、上述の河合の報告によれば、クラスタリングの精度は６０％ほどで、実用化にはほど遠い。また、カテゴライゼーションは、汎用的な分類にとどまり

、元の文書群の特徴を反映しない。

【００１６】

本発明の課題は、情報処理装置に蓄えられた大量の文書群を、その特徴に従って高い精度で自動的に分類する文書整理装置およびその方法を提供することである。

【００１７】

【課題を解決するための手段】

図１は、本発明の文書整理装置の原理図である。図１の文書整理装置は、関係抽出手段１、生成手段２、および出力手段３を備え、文書群をキーワードに基づいて整理する。

【００１８】

関係抽出手段１は、与えられたキーワードからキーワード間の階層関係４（実線）と連想関係５（破線）を抽出する。生成手段２は、階層関係４と連想関係５をリンクとして用いて、上記文書群にアクセスするためのディレクトリ情報を生成する。そして、出力手段３は、そのディレクトリ情報を出力する。

【００１９】

関係抽出手段１により抽出される階層関係４は、キーワードの間の概念上の上下関係を表し、連想関係５は、階層関係ほど緊密ではないが、一方のキーワードから他方のキーワードが連想されるような比較的緩やかな関係を表す。この連想関係５により、階層関係４を持たないキーワード同士を関係付けることが可能になる。

【００２０】

生成手段２は、階層関係４だけでなく連想関係５もリンクとして用いて、文書に付加されたキーワード間の関係を表すディレクトリ情報を生成し、文書を分類する。そして、出力手段３は、利用者が文書群にアクセスできるように、ディレクトリ情報を文書群の索引として提示する。

【００２１】

文書群の分類結果を表すディレクトリ情報に連想関係５をリンクとして付加することにより、階層関係４だけでは得られなかったフレキシブルなアクセスが可

能となる。したがって、文書群の特徴を分類に反映させることがより容易になり、高精度の分類結果が自動的に得られる。

【００２２】

例えば、図１の関係抽出手段１は、後述する図２のキーワード関係抽出器４２に対応し、生成手段２は、図２のディレクトリファイル生成器４３に対応し、出力手段３は、図２の表示装置１４およびディレクトリアクセス部４４に対応する。

【００２３】

【発明の実施の形態】

以下、図面を参照しながら、本発明の実施の形態を詳細に説明する。

本発明においては、管理に要する人手を最小限にするために、管理者は、キーワードの階層関係の極一部だけを辞書や人手で文書整理装置に与える。文書整理装置は、それを元に統計処理および文字列処理を行って、キーワード間の関係を自動的に取り出し、ディレクトリファイルを自動生成する。また、既存の分類方法の精度上の限界を克服するために、文書の分類（階層関係）にはこだわらず、利用者が文書にアプローチするための多様なリンクを提供することをディレクトリの主眼とする。この文書整理装置の動作の概要は次のようになる。

【００２４】

１．まず、入力文書にはその内容を表すキーワードが付与されているものとする。キーワード付与の方法としては、本発明とは独立に、任意の方法を用いることができる。例えば、キーワードを人手で付けてもよいし、文書内容からキーワード抽出技術により自動的に取り出してもよい。

【００２５】

２．次に、管理者は、ディレクトリのトップとなるキーワード列、キーワードの同値関係、キーワードの階層（上下）関係、および不要キーワードのリストを、明示的に文書整理装置に与える。

【００２６】

３．文書整理装置は、管理者が明示した不要キーワードのリストと、キーワード変換ルールに基づいて、文書に付与されたキーワードを整える（データクリー

ニング）。

【００２７】

４．文書整理装置は、各キーワードを含む文書集合を計算し、管理者が同意語として明示した同値関係にある２つ以上のキーワードの文書集合をマージする。

５．文書整理装置は、任意の２つのキーワードに対して、それらの文書集合の関係をもとに、キーワード間の関連性（同値関係／階層関係／連想関係）を計算する。

【００２８】

６．文書整理装置は、管理者が明示した階層関係および計算により得られた階層関係を元に、管理者が与えたトップキーワードから特定のキーワードまでの最短距離のキーワード列を、そのキーワード（およびそれを有する文書）へのパスとして計算する。

【００２９】

７．文書整理装置は、まだパスの付いていないキーワードについて、そのキーワードから最も少ない上位関連語列により辿りつけるパスの付いたキーワードを探索する。そして、得られたキーワードのパスに上位関連語列を付加したものを、元のキーワードのパスとする。

【００３０】

８．それでもパスが付けられないキーワードについては、文書整理装置は、トップキーワードに“その他”を設け、その下位語として登録する。

９．文書整理装置は、単語読み付加器を用いて、キーワードに読みを与える。

【００３１】

１０．文書整理装置は、管理者が明示したキーワード間の階層関係、自動的に得られたキーワード間の関係、自動的に計算されたパス、キーワードの読み、およびキーワード分割装置（部分文字列切り出し器）を用いて、ディレクトリのハイパーテキストを作成する。

【００３２】

１１．利用者は、作成されたディレクトリを文書群の索引として用いて、次のような操作を行う。

・トップレベルから下位関連語を用いて文書の検索範囲を絞り込む。

【００３３】

・上位関連語を用いて検索範囲を広げる。

・パスで現在位置を確認し、上位に飛ぶ。

・連想関係のキーワードを用いて別の概念に飛ぶ。

【００３４】

・５０音・アルファベット順索引を用いてキーワードを探す。

図２は、このような文書整理装置の構成図である。図２の文書整理装置は、処理装置１１、二次記憶装置１２、入力装置１３、および表示装置１４を備える。処理装置１１は、例えば、ＣＰＵ（中央処理装置）とメモリを含み、入力装置１３は、キーボード、マウス等に対応し、表示装置１４は、ディスプレイ等に対応する。

【００３５】

また、二次記憶装置１２は、電子化された文書群のデータ２１と、ディレクトリ管理者が辞書等を用いて作成／管理する管理ファイル２２を格納する。データ２１の各文書には、それぞれ、複数のキーワードが付加されている。また、管理ファイル２２には、同意語および不要語の集合３１と、キーワードの概念上の階層関係を表すデータ３２と、ディレクトリのトップとなるキーワードの集合３３が含まれている。

【００３６】

ここで、同意語とは、“コンピュータ”と“計算機”のように、互いに同意語関係にあるキーワードの集合を表し、不要語とは、差別用語のように、ディレクトリに用いたくないキーワードの集合を表す。また、階層関係としては、例えば、“コンピュータ”を上位語（上位関連語）として、“ハードウェア”、“ソフトウェア”等の下位語（下位関連語）を定義する情報が用いられる。

【００３７】

また、処理装置１１は、キーワード整形器４１、キーワード関係抽出器４２、ディレクトリファイル生成器４３、ディレクトリアクセス部４４、および検索部４５を含む。これらは、例えば、プログラムにより記述されたソフトウェアコン

ポーネントに対応し、処理装置１１の特定のプログラムコードセグメントに格納される。

【００３８】

キーワード整形器４１は、いわゆるデータクリーニングを行い、文書単語対５１と文書メタ情報５２を出力する。ここでは、文書データ２１の中のゴミを取ったり、キーワードを統一したりする処理が行われる。文書単語対５１は、整形後の文書とキーワードの対であり、文書メタ情報１０２は、文書の諸々の情報（ＵＲＬ、タイトル等）やキーワードの読み等の情報である。

【００３９】

キーワード関係抽出器４２は、データ３２、５１から、キーワード間の階層関係５３、同値関係５４、および連想関係５５を計算する。これらの関係は、いずれも２つ以上のキーワード間になんらかの関連性があることを表している。

【００４０】

階層関係５３は、キーワードの間の概念上の上下関係を表すデータであり、同値関係５４は、キーワードの間の概念上の同位関係を表すデータである。例えば、同意語の集合に含まれるキーワードは、互いに同値関係にある。また、連想関係５５は、階層関係も同値関係も持たないキーワード間において、一方のキーワードから他方のキーワードが連想されるような関係を表すデータである。

【００４１】

ディレクトリファイル生成器４３は、データ３３、５２、５３、５４、５５から、ハイパーテキスト形式のディレクトリファイル５６を生成し、ディレクトリアクセス部４４は、入力装置１３から入力される利用者の指示に従って、ディレクトリファイル５６の内容を表示装置１４に表示したり、検索部４５を起動したりする。

【００４２】

検索部４５は、利用者の指示に従って、ディレクトリファイル５６や文書データ２１の内容を全文検索する。文書データ２１の全文検索では、文書の本文に含まれる任意の文字／単語列の検索が行われる。

【００４３】

次に、図３から図５までを参照しながら、各種データのデータ構造について説明する。

図２の管理ファイル２２に含まれるデータ３１、３２、３３は、例えば、図３に示すような３つのテーブルに格納される。図３の同意語／不要語テーブル３１は、［代表語ＩＤ，同意語または不要語のＩＤの集合］の組合せを表す。例えば、代表語ＩＤ＝００５のキーワードは、ＩＤ＝０１０，０２１，０２２の他のキーワードと同意語の関係にあり、代表語ＩＤが空（ｎｉｌ）のＩＤ集合（ＩＤ＝０７７，０８２）は、不要語である。

【００４４】

また、階層関係テーブル３２は、［上位語ＩＤ，下位語ＩＤの集合］の組合せを表す。例えば、上位語ＩＤ＝００２のキーワードは、階層関係における下位語としてＩＤ＝００４，００８のキーワードを有する。また、トップキーワードＩＤリスト３３は、トップディレクトリとして用いられるキーワードのＩＤの列（ＩＤ＝００１，００８，．．．）を表す。

【００４５】

また、図２の文書単語対５１、階層関係５３、同値関係５４、連想関係５５は、例えば、図４に示すような形式の２つのテーブルに格納される。図４の文書単語対テーブル６１は、

［文書ＩＤ，キーワードＩＤの集合］

の組合せを格納し、キーワードテーブル６２は、

［キーワードＩＤ（ＫＩＤ），キーワード，読み，上位語ＩＤの集合（ＵＰ），下位語ＩＤの集合（ＤＯＷＮ），連想語ＩＤの集合（Ｒｅｌ），同値関係にあるキーワードＩＤの集合（Ｅｑ），パス，新規語フラグ（ｎｅｗ）］

の組合せを格納する。

【００４６】

ここで、上位語ＩＤの集合と下位語ＩＤの集合は階層関係５３に対応し、連想語ＩＤの集合は連想関係５５に対応し、同値関係にあるキーワードＩＤの集合は同値関係５４に対応する。また、新規語フラグは、キーワードが新規語に対応する場合は“１”となり、そうでない場合は“０”となる。

【００４７】

例えば、文書単語対テーブル６１の文書ＩＤ＝０００の文書には、ＩＤ＝０００，０８１のキーワードが付加されており、ＩＤ＝０００のキーワードは、キーワードテーブル６２の“スポーツ”に相当する。

【００４８】

このキーワードの読みとしては、“すぽーつ”が登録されており、その上位語としては、ＩＤ＝００８，０２２の２つのキーワードが登録されている。また、その下位語としては、ＩＤ＝０２５のキーワードが登録されており、連想語としては、ＩＤ＝０３８，０８７の２つのキーワードが登録されている。この段階では、同値関係にあるキーワードとパスは登録されておらず、新規語フラグ“１”は、“スポーツ”が新規語であることを表している。

【００４９】

また、図２の文書メタ情報５２は、図５に示すような文書情報テーブルの形式で格納される。この文書情報テーブル５２は、

［文書ＩＤ，タイトル，説明，更新日時，一次情報へのリンク］

の組合せを格納する。

【００５０】

ここで、説明は、文書の要約または最初の数行の記述に対応し、更新日時は、文書が最後に更新された日時を表し、一次情報へのリンクは、インターネット上の対応するＵＲＬ等を表す。

【００５１】

例えば、文書ＩＤ＝００１の文書のタイトルは“ニュースの読み方”であり、その説明は“comp,fj をmuleから．．．”であり、更新日時は“１９９８年２月１０日１５時３８分”であり、一次情報へのリンクは“ｈｔｔｐ：／／ｗｗｗ．ｘｘｘｘ（ＵＲＬ）”である。

【００５２】

次に、図６および図７を参照しながらキーワード整形器４１の処理を詳細に説明する。

図６は、キーワード整形器４１の構成図である。図６のキーワード整形器４１

は、サブコンポーネントとして、キーワード統一器７１および単語読み付加器７２を含む。キーワード統一器７１は、与えられたキーワードの文字コードを統一し、単語読み付加器７２は、与えられた単語の読みを生成する。

【００５３】

このキーワード整形器４１は、図７に示すような処理を行って、文書単語対５１と文書情報テーブル５２を生成する。文書単語対５１は、図４に示したように、文書単語対テーブル６１とキーワードテーブル６２から成っている。

【００５４】

まず、キーワード統一器７１は、文書データ２１の文書に付加されたすべてのキーワードについて、文字を統一する（ステップＳ１０）。ここでは、キーワードに含まれる漢字コードを、例えば、ＥＵＣコードのような特定のコードに統一したり、半角のカタカナ／英数字を全角のカタカナ／英数字に統一したり、キーワード中の記号や空白を除去したりするような、文字単位でのキーワード統一処理が行われる。

【００５５】

次に、キーワード整形器４１は、１つの文書のキーワード情報のうち、キーワードＩＤとキーワードをキーワードテーブル６２に追加し、読みと新規語フラグを付加する（ステップＳ１１）。

【００５６】

このとき、日本語のキーワードの読みについては、単語読み付加器７２が生成した読みを設定し、英語のキーワードの読みについては、キーワード自身と同一に設定する。また、追加したキーワードが旧キーワードの集合７３に登録されていれば、新規語フラグを“０”に設定し、旧キーワードの集合７３に登録されていなければ、新規語フラグを“１”に設定する。旧キーワードの集合７３は、前回のディレクトリ生成時に生成されたキーワードテーブル６２のキーワードの集合を表す。

【００５７】

次に、キーワード整形器４１は、文書ＩＤおよびキーワードＩＤを、文書メタ情報５２およびキーワードテーブル６２へのリンクとして文書単語対テーブル６

１に追加する（ステップＳ１２）。このとき、対応する文書に関するデータを文書情報テーブル５２に追加しておく。

【００５８】

次に、文書データ２１の中に未処理の文書があるかどうかを判定し（ステップＳ１３）、そのような文書があれば、ステップＳ１１以降の処理を繰り返す。そして、すべての文書について登録が完了すると、次に、同意語／不要語テーブル３１の情報をキーワードテーブル６２に追加する（ステップＳ１４）。

【００５９】

ここでは、テーブル３１の代表語ＩＤに対応するキーワードのフィールドＥｑに、その同意語として定義されたＩＤを追加し、不要語として定義されたＩＤをキーワードテーブル６２のエントリからすべて消去する。また、同時に、文書単語対テーブル６１から不要語のＩＤをすべて消去する。これにより、管理者が指定した同意語がキーワードテーブル６２に登録され、不要語が文書単語対テーブル６１およびキーワードテーブル６２から消去される。

【００６０】

次に、キーワードテーブル６２のエントリとして、“コンピューター”と“コンピュータ”のように、語尾の文字“ー”が有るキーワードとそれが無いキーワードとが併存する場合には、それらを統一する（ステップＳ１５）。ここでは、例えば、“コンピューター”のフィールドＥｑに“コンピュータ”のキーワードＩＤを登録することにより、これらの２つのキーワードを縮退させることができる。

【００６１】

また、“コンピュータグラフィックス”と“コンピュータ”のように、あるキーワードとそれに含まれる他のキーワードに関しても統一し、それらのエントリが併存する場合には、一方のフィールドＥｑに他方のキーワードＩＤを登録する。

【００６２】

次に、こうして生成されたキーワードテーブル６２のキーワードのリストを、旧キーワードの集合７３として保存し（ステップＳ１６）、処理を終了する。旧

キーワードの集合７３は、次回のディレクトリ生成時に、前述のステップＳ１１の処理で参照される。

【００６３】

次に、図８から図１５までを参照しながらキーワード関係抽出器４２の処理を詳細に説明する。

図８は、キーワード関係抽出器４２の構成図である。図８のキーワード関係抽出器４２は、部分文字列切り出し器８１、相関ルール抽出器８２、ルール評価器８３、およびマージャ８４を含み、文書単語対５１（図４の文書単語対テーブル６１とキーワードテーブル６２に対応）と管理者が与えた階層関係３２から、階層関係５３、同値関係５４、および連想関係５５を生成する。

【００６４】

部分文字列切り出し器８１は、複合語句のキーワードを適当な文字列に分割し、部分語関係のデータ８５を生成する。データ８５は、“情報検索”と“情報”／“検索”のように、キーワード間で一方が他方を含むような関係を表す。

【００６５】

相関ルール抽出器８２は、キーワードの出現頻度から、キーワード間の関係を表す相関ルール８６を抽出する。ルール評価器８３は、相関ルール８６を評価し、それを階層関係、同値関係５４、および連想関係５５に分割する。マージャ８４は、階層関係３２、部分語関係８５、およびルール評価器８３が生成した階層関係を合わせて、階層関係５３を生成する。

【００６６】

図９は、キーワード関係抽出器４２の処理のフローチャートである。まず、相関ルール抽出器８２は、文書単語対テーブル６１とキーワードテーブル６２から相関ルール８６を抽出し、それを相関ルールの集合ＡＲとする（ステップＳ２１）。

【００６７】

次に、ルール評価器８３は、ＡＲに基づいてルール評価を行い、それを上位語関係ｕｐ１、下位語関係ｄｏｗｎ１、同値関係ｅｑ１、および連想関係ｒｗ１に分割する（ステップＳ２２）。２つのキーワードｗ１、ｗ２について、ｕｐ１お

よびｄｏｗｎ１は、“ｗ１＞ｗ２”（ｗ１はｗ２の上位語、またはｗ２はｗ１の下位語）のように記述され、ｅｑ１は“ｗ１＝ｗ２”のように記述され、ｒｗ１は“ｗ１～ｗ２”のように記述される。

【００６８】

次に、ルール評価器８３は、同値関係ｅｑ１をキーワードテーブル６２のフィールドＥｑに登録する（ステップＳ２３）。ｅｑ１は、“ｗ１＝ｗ２”のようなキーワード間の関係の集合である。そこで、キーワードｗ１のフィールドＥｑのエントリに、キーワードｗ２のＩＤを追加する。

【００６９】

次に、ルール評価器８３は、連想関係ｒｗ１をキーワードテーブル６２のフィールドＲｅ１に登録する（ステップＳ２４）。ｒｗ１は、“ｗ１～ｗ２”のようなキーワード間の関係の集合である。そこで、キーワードｗ１のフィールドＲｅ１のエントリに、キーワードｗ２のＩＤを追加する。

【００７０】

次に、部分文字列切り出し器８１は、キーワードテーブル６２のキーワードの中から部分文字列を取り出し、それらの包含関係を表す部分語関係８５を生成する（ステップＳ２５）。そして、それを部分語関係の集合ｓｒ１とおく。ｓｒ１には、例えば、“コンピュータ”と“グラフィックス”が“コンピュータグラフィックス”の部分文字列であることが登録される。

【００７１】

次に、マージャ８４は、部分語関係ｓｒ１を参照し、キーワードｋｗが部分文字列w_1，w_2，．．．，w_nを含んでいる場合、w_iがｋｗの上位語（ｉ＝１，２，．．．，ｎ）であるという階層関係に変換し、それを“w_i＞ｋｗ”と記述する（ステップＳ２６）。こうして得られた階層関係の集合をｓｒ２とする。

【００７２】

次に、マージャ８４は、管理者が定義した階層関係３２にｓｒ２をマージし、その結果得られた階層関係をＨＲとする（ステップＳ２７）。ここで、ある階層関係Ｓ１に他の階層関係Ｓ２をマージするとは、Ｓ２の要素のうち、Ｓ１の各要素と矛盾しないものを新たにＳ１に加える操作を意味する。例えば、Ｓ２の要素

“ｗ１＞ｗ２”がＳ１になく、“ｗ２＞ｗ１”という要素もＳ１にないとき、“ｗ１＞ｗ２”がＳ１に加えられる。

【００７３】

ここでは、階層関係３２の方がｓｒ２より優先されるため、ｓｒ２の要素のうち、階層関係３２の要素と矛盾するものは、階層関係３２に加えられない。例えば、階層関係３２が、

｛コンピュータ＞ソフトウェア，コンピュータ＞ハードウェア，ソフトウェア＞グループウェア，グローバルネットワーク＞ネットワーク｝

であり、ｓｒ２が、

｛コンピュータ＞パーソナルコンピュータ，ネットワーク＞グローバルネットワーク｝

である場合、階層関係３２にｓｒ２をマージした結果は、

｛コンピュータ＞ソフトウェア，コンピュータ＞ハードウェア，コンピュータ＞パーソナルコンピュータ，ソフトウェア＞グループウェア，グローバルネットワーク＞ネットワーク｝

となる。

【００７４】

次に、マージャ８４は、ステップＳ２２で得られたｄｏｗｎ１をＨＲにマージし、その結果を改めてＨＲとする（ステップＳ２８）。これにより、文字列の包含関係により得られた階層関係ｓｒ２が、ルール評価により得られた階層関係ｄｏｗｎ１よりも優先されることになる。

【００７５】

次に、マージャ８４は、ＨＲの各要素“ｗ１＞ｗ２”について、キーワードテーブル６２のキーワードｗ１のレコードのフィールドＤＯＷＮに、キーワードｗ２のＩＤを追加する（ステップＳ２９）。

【００７６】

次に、マージャ８４は、ステップＳ２２で得られたｕｐ１をＨＲにマージし、その結果を改めてＨＲとする（ステップＳ３０）。これにより、上位語関係もＨＲに含められる。

【００７７】

次に、マージャ８４は、ＨＲの各要素“ｗ１＞ｗ２”について、キーワードテーブル６２のキーワードｗ２のレコードのフィールドＵＰに、キーワードｗ１のＩＤを追加して（ステップＳ３１）、処理を終了する。

【００７８】

図１０は、図９のステップＳ２１で行われる相関ルール抽出処理と、ステップＳ２２で行われるルール評価処理のフローチャートである。ここで、与えられる入力データは、キーワード整形済みの文書単語対テーブル６１とキーワードテーブル６２である。

【００７９】

まず、相関ルール抽出器８２は、キーワードテーブル６２の同値関係を参照して、文書毎に同値関係を解消したキーワードの集合を生成し、データマイニングの相関ルール抽出アルゴリズムを利用して、相関ルールを抽出する（ステップＳ４１）。

【００８０】

ここで、同値関係を解消したキーワードの集合とは、文書単語対テーブル６１に文書と対応して登録されているキーワードに、それらのキーワードのフィールドＥｑに登録されたキーワードを加え、相異なるものだけを残したキーワードの集合を意味する。

【００８１】

また、相関ルール抽出アルゴリズムによれば、キーワード対（Ｈ，Ｂ）に関する統計情報に基づき、文書とキーワード集合の対からＨ→Ｂという形式のルールが抽出される。ここで、Ｈ→Ｂというルールは、キーワード対（Ｈ，Ｂ）の共起出現頻度を表すサポートｓｕｐ（Ｈ→Ｂ）および確信度ｃｏｎｆ（Ｈ→Ｂ）により特徴付けられる。これらの値は、次式により定義される。

ｓｕｐ（Ｈ→Ｂ）

＝（Ｈ　ａｎｄ　Ｂを有する文書数）／全文書数

ｃｏｎｆ（Ｈ→Ｂ）

＝（Ｈ　ａｎｄ　Ｂを有する文書数）／（Ｈを有する文書数）

ステップＳ４１では、ＳｕｐおよびＣｆを適当な閾値として、すべてのキーワードH、Ｂの組合せのうち、ｓｕｐ（Ｈ→Ｂ）≧Ｓｕｐおよびｃｏｎｆ（Ｈ→Ｂ）≧Ｃｆを満たすような、ＨとＢの組合せをルールとして取り出す。

【００８２】

次に、ルール評価器８３は、このようにして取り出されたルールの集合を分割して、キーワード間の階層関係、同値関係、および連想関係を取り出す。これにより、各ルールに含まれるキーワード対の関係が自動推定される。

【００８３】

ここでは、ｋｗとw_i（ｗ＿ｉ）をルールに含まれるキーワードＩＤとして、ｃｏｎｆ（ｋｗ→w_i）の値をＸ軸に取り、ｃｏｎｆ（w_i→ｋｗ）の値をＹ軸に取って、図１１に示すような確信度のＸＹ平面を考える。

【００８４】

このとき、ルールの集合に含まれるｋｗとw_iのすべての組合せについて、Ｃｆ≦ｃｏｎｆ（ｋｗ→w_i）≦１、Ｃｆ≦ｃｏｎｆ（w_i→ｋｗ）≦１が成り立つ。したがって、各w_iについて、点（Ｘ，Ｙ）＝（ｃｏｎｆ（ｋｗ→w_i），ｃｏｎｆ（w_i→ｋｗ））をＸＹ平面上にプロットすると、×印で示されるように、Ｘ軸、Ｙ軸、直線Ｘ＝Ｃｆ、直線Ｘ＝１、直線Ｙ＝Ｃｆ、および直線Ｙ＝１で囲まれた領域内に点が分布することになる。

【００８５】

そこで、適当な閾値をＴｘ、Ｔｙとして、直線Ｘ＝ＴｘとＹ＝Ｔｙによりこの領域を４つの部分領域９１、９２、９３、９４に分割し、それぞれの部分領域に含まれるw_iに、ｋｗとの間の階層関係、同値関係、および連想関係のいずれかを付与する。

ルール評価器８３は、まず、変数ｋｗを０００（キーワードＩＤの最小値）とおく（ステップＳ４２）。そして、ｋｗを左辺または右辺に持つルールを要素とする集合をＳとおき、変数Ｔｘ、Ｔｙの初期値をそれぞれ１に設定する（ステップＳ４３）。

【００８６】

次に、Ｓの要素のうち、ｃｏｎｆ（ｋｗ→w_i）＞Ｔｘとなるようなルールに

含まれるw_iの数ｓｘ（Ｓ，Ｔｘ）を求め、それが一定値min_x（ｍｉｎ＿ｘ）を越えるまで、Ｔｘの値を徐々に下げていく（ステップＳ４４～Ｓ４７）。この処理により、図１１の領域９１、９２を合わせた領域のキーワード数がmin_x以上になるように、Ｔｘが決められる。ただし、Ｔｘは最小確信度Ｃｆより下には下げないものとする。

【００８７】

ここでは、まず、ｓｘ（Ｓ，Ｔｘ）をmin_xと比較し（ステップＳ４４）、ｓｘ（Ｓ，Ｔｘ）がmin_x以下であれば、Ｔｘの値を０．１だけ下げる（ステップＳ４５）。そして、ＴｘをＣｆと比較し（ステップＳ４６）、ＴｘがＣｆより大きければ、ステップＳ４４以降の処理を繰り返す。そして、ステップＳ４４においてｓｘ（Ｓ，Ｔｘ）がmin_xを越えれば、ステップＳ４８以降の処理に移る。また、ステップＳ４６においてＴｘがＣｆ以下となった場合は、Ｔｘ＝Ｃｆとおき（ステップＳ４７）、ステップＳ４８以降の処理に移る。

【００８８】

次に、Ｓの要素のうち、ｃｏｎｆ（w_i→ｋｗ）＞Ｔｙとなるようなルールに含まれるw_iの数ｓｙ（Ｓ，Ｔｙ）を求め、その数が一定値min_y（ｍｉｎ＿ｙ）を越えるまで、Ｔｙの値を徐々に下げていく（ステップＳ４８～Ｓ５１）。この処理により、図１１の領域９１、９３を合わせた領域のキーワード数がmin_y以上になるように、Ｔｙが決められる。ただし、Ｔｙは最小確信度Ｃｆより下には下げないものとする。

【００８９】

ここでは、まず、ｓｙ（Ｓ，Ｔｙ）をmin_yと比較し（ステップＳ４８）、ｓｙ（Ｓ，Ｔｙ）がmin_y以下であれば、Ｔｙの値を０．１だけ下げる（ステップＳ４９）。そして、ＴｙをＣｆと比較し（ステップＳ５０）、ＴｙがＣｆより大きければ、ステップＳ４８以降の処理を繰り返す。そして、ステップＳ４８においてｓｙ（Ｓ，Ｔｙ）がmin_yを越えれば、ステップＳ５２以降の処理に移る。また、ステップＳ５０においてＴｙがＣｆ以下となった場合は、Ｔｙ＝Ｃｆとおき（ステップＳ５１）、ステップＳ５２以降の処理に移る。

【００９０】

次に、図１１に示したように、ｋｗに関連する確信度を表す点（Ｘ，Ｙ）＝（ｃｏｎｆ（ｋｗ→w_i），ｃｏｎｆ（w_i→ｋｗ））をＸＹ平面上にプロットする（ステップＳ５２）。

【００９１】

このとき、右上の矩形領域９１に属する各点は、Ｘ、Ｙの値が共に１に近く（大きく）、図１２に示すように、w_iの文書集合（w_iをキーワードとして有する文書の集合）はｋｗの文書集合とほぼ重なると考えられる。このため、w_iはｋｗと同値関係にあるものとみなされる。

【００９２】

また、右下の矩形領域９２に属する各点は、Ｘの値が１に近くＹの値が小さいので、図１３に示すように、w_iの文書集合はｋｗの文書集合をほぼ含んでいると考えられる。このため、w_iはｋｗの上位語であるものとみなされる。

【００９３】

また、左上の矩形領域９３に属する各点は、Ｙの値が１に近くＸの値が小さいので、図１４に示すように、w_iの文書集合はｋｗの文書集合にほぼ含まれると考えられる。このため、w_iはｋｗの下位語であるものとみなされる。

【００９４】

また、左下の領域９４に属する各点は、Ｘ，Ｙの値が共に小さいが最小確信度Ｃｆ以上であるので、w_iの文書集合とｋｗの文書集合は、上述のような関係にはないが、図１５に示すように、なんらかの関連性を持つと考えられる。このため、w_iはｋｗの連想語であるものとみなされる。

【００９５】

そこで、Ｓの要素を対応する４つのグループに分割し、領域９１のw_iについては同値関係“ｋｗ＝w_i”をｅｑ１に加え、領域９４のw_iについては連想関係“ｋｗ～w_i”をｒｗ１に加える。また、領域９２のw_iについては上位語関係“w_i＞ｋｗ”をｕｐ１に加え、領域９３のw_iについては、下位語関係“ｋｗ＞w_i”をｄｏｗｎ１に加える。こうして、ｋｗに関するすべてのキーワード関係が抽出される。同値関係は、上位語関係および下位語関係とともに、広義の階層関係に属するとも考えられる。

【００９６】

次に、ルール評価器８３は、ｋｗに１を加算して（ステップＳ５３）、ｋｗをキーワードＩＤの最大値max_{kw}（ｍａｘ＿ｋｗ）と比較する（ステップＳ５４）。そして、ｋｗがmax_{kw}を越えていなければ、ステップＳ４３以降の処理を繰り返し、ｋｗがmax_{kw}を越えると、処理を終了する。

【００９７】

次に、図１６から図１９までを参照しながらディレクトリファイル生成器４３の処理を詳細に説明する。

ディレクトリファイル生成器４３は、キーワード関係抽出器４２が生成した階層関係５３、同値関係５４、連想関係５５、管理者が与えたトップキーワードＩＤリスト３３、およびキーワード整形器４１が生成した文書メタ情報５２から、ディレクトリファイル５６を生成する。

【００９８】

ディレクトリファイル５６は、図１６に示すような３種類のハイパーテキストファイル１０１、１０２、１０３から成り、各ファイルの間には互いにリンクが張られている。

【００９９】

図１６において、ディレクトリトップファイル１０１は、ディレクトリの入口に対応するファイルであり、１つだけ設けられる。このファイル１０１には、キーワード検索の入力窓１０４、トップキーワード１０５（ＫＬ）、および５０音・アルファベット順索引１０６が含まれている。トップキーワード１０５の各キーワードＫＬからは、ディレクトリ中間ファイル１０３へリンクが張られており、５０音・アルファベット順索引１０６の各文字からは、５０音・アルファベット順索引中間ファイル１０２へリンクが張られている。

【０１００】

５０音・アルファベット順索引中間ファイル１０２は、“あ行”、“か行”等キーワードの読みにより複数のファイルに分割される。各ファイルのキーワードＫＬからは、ディレクトリ中間ファイル１０３にリンクが張られている。

【０１０１】

ディレクトリ中間ファイル１０３は、キーワード毎に設けられ、ヘッダ１０７、パス１０８、上位関連語１０９、サブカテゴリ１１０、および文書リスト１１１の各部分から構成される。

【０１０２】

ヘッダ１０７には、ファイル１０３のタイトルとなるキーワードと、その同意語リストが記述され、パス１０８には、トップキーワードからそのキーワードまでの経路の１つがキーワード列として記述される。パス１０８の各キーワードＫＬからは、そのキーワードのディレクトリ中間ファイル１０３へリンクが張られている。

【０１０３】

また、上位関連語１０９には、上位キーワード列が記述され、サブカテゴリ１１０には、下位キーワード列が記述される。各キーワードＫＬからは、そのキーワードのディレクトリ中間ファイル１０３へリンクが張られている。

【０１０４】

文書リスト１１１には、そのキーワードと関連付けられた各文書のタイトルと内容が記述される。タイトルからは、文書の一次情報（ＵＲＬ等）へのリンクが張られている。また、各文書に付加されたキーワードのうち、そのファイルのキーワードと連想関係にあるものが連想語として記述され、各連想語からはそのディレクトリ中間ファイル１０３へリンクが張られている。検索窓１１２は、文書の内容を検索する際の入力窓である。

【０１０５】

なお、図１６において、★印の付いたキーワードは新規語に対応する。この印は、図４のキーワードテーブル６２のフィールドｎｅｗに“１”が設定されたキーワードに対して付加され、それらが新規語であることを強調している。

【０１０６】

図１７は、ディレクトリファイル生成器４３の処理のフローチャートである。ディレクトリファイル生成器４３は、まず、ディレクトリトップファイル１０１を生成し、データ３３のキーワードを領域１０５に記述する（ステップＳ６１）。また、５０音・アルファベット順索引中間ファイル１０２へのリンクを領域１

０６に記述する。

【０１０７】

次に、５０音・アルファベット順索引中間ファイル１０２を生成し、キーワードテーブル６２に登録された読みを取得して、同一の読みで始まるキーワードを１つのファイル１０２にまとめる（ステップＳ６２）。

【０１０８】

次に、トップキーワードから各キーワードまでの最短パスを計算し、それをキーワードテーブル６２のパスのフィールドに登録する（ステップＳ６３）。そして、キーワード毎にディレクトリ中間ファイル１０３を生成して（ステップＳ６４）、処理を終了する。

【０１０９】

ステップＳ６４では、キーワードテーブル６２のフィールドＥｑに登録されたキーワードを領域１０７に記述し、計算された最短パスを領域１０８に記述し、キーワードテーブル６２のフィールドＵＰに登録されたキーワードを領域１０９に記述し、フィールドＤＯＷＮに登録されたキーワードを領域１１０に記述する。

【０１１０】

また、文書単語対テーブル６１から、ファイル１０３のタイトルキーワードが付加された文書のＩＤを取得し、文書情報テーブル５２から、そのタイトル、説明、および一次情報へのリンクを取得する。そして、それらを領域１１１に記述する。このとき、キーワードテーブル６２のフィールドＲｅｌに登録されたキーワードを、連想語として記述する。したがって、ファイル１０３の生成において、パス１０８だけが新規に作成される情報である。

【０１１１】

ここで、トップキーワードからあるキーワードまでのパスとしては、トップから階層関係だけを辿って到達するものだけでなく、それに連想関係を加えて到達するものも存在する。これは、前述のような階層関係の設定方法では、トップから階層関係だけで必ずしもすべてのキーワードに到達できるという保証がないためである。

【０１１２】

そもそも、パスはハイパーテキストにおいて利用者が迷子にならないための仕組みである。本実施形態においては、利用者は、パスを逆に辿ることで、あるキーワードからトップキーワードへ到達することができる。

【０１１３】

図１８および図１９は、図１７のステップＳ６３におけるパス生成処理のフローチャートである。ここでは、最初に階層関係だけでパスの生成を試み、パスが生成されなかったキーワードについては、連想関係も加えてパス生成を試みる。それでもトップと結びつけられないキーワードについては、便宜上、トップの直下に“その他”というカテゴリを生成し、そこに直接結びつける。

【０１１４】

ディレクトリファイル生成器４３は、まず、トップキーワードＩＤリスト３３の各キーワードについて、キーワードテーブル６２のパスのフィールドに“ｔｏｐ”を登録し（図１８、ステップＳ７１）、それらのトップキーワードのリストをＳ１とする（ステップＳ７２）。

【０１１５】

次に、階層関係だけを辿ってパスを設定する（ステップＳ７３～Ｓ８０）。ここでは、幅優先探索を行っており、各時点での最もパスが長いキーワードがＳ１に入っている。

【０１１６】

ディレクトリファイル生成器４３は、まず、Ｓ１が空かどうかを判定し（ステップＳ７３）、それが空でなければ、Ｓ１からキーワードｗを取り出す（ステップＳ７４）。そして、キーワードテーブル６２において、キーワードｗのフィールドＤＯＷＮに登録されたキーワード集合をＳ２とする（ステップＳ７５）。

【０１１７】

次に、Ｓ２が空かどうかを判定し（ステップＳ７６）、それが空でなければ、Ｓ２からキーワードｕを取り出し（ステップＳ７７）、そのパスのフィールドが空かどうかを調べる（ステップＳ７８）。

【０１１８】

まだ、キーワードｕのパスが設定されていなければ、ｗのパス＋ｗをパスとして設定し、キーワードｕをＳ３に加えて（ステップＳ７９）、ステップＳ７６以降の処理を繰り返す。Ｓ３は、パスが設定されたキーワードの集合を表し、最初は空に設定されている。ステップＳ７８においてキーワードｕのパスが既に設定されていれば、そのままステップＳ７６以降の処理を繰り返す。

【０１１９】

そして、ステップＳ７６においてＳ２が空になると、Ｓ３を改めてＳ１とおき、Ｓ３を空に設定して（ステップＳ８０）、ステップＳ７３以降の処理を繰り返す。これにより、キーワードの木構造のトップから下位に向かって、各ノードにパスが設定されていく。

【０１２０】

ステップＳ７３においてＳ１が空になると、次に、キーワードテーブル６２においてまだパスの設定されていないキーワードに対して、階層関係および連想関係を辿ってパスを設定する（ステップＳ８１～Ｓ８９）。

【０１２１】

ディレクトリファイル生成器４３は、まず、その時点でまだパスが設定されていないキーワードの集合をＳ４とし、Ｓ６を空に設定する（ステップＳ８１）。次に、Ｓ４が空かどうかを判定し（ステップＳ８２）、それが空でなければ、Ｓ４からキーワードｖを取り出す。そして、キーワードテーブル６２において、キーワードｖのフィールドＵＰ、ＤＯＷＮ、Ｒｅｌを合わせたキーワードの集合をＳ５とする（ステップＳ８３）。

【０１２２】

次に、Ｓ５のキーワードでパスが設定されているもののうち、最短パスのキーワードをｗとする（ステップＳ８４）。Ｓ５のキーワードのパスがすべて空の場合には、ｗも空となる。

【０１２３】

次に、キーワードｗが空かどうかを判定し（ステップＳ８５）、それが空でなければ、ｗのパス＋ｗをキーワードｖのパスとして設定し、Ｓ６にキーワードｖを加える（ステップＳ８６）。そして、ステップＳ８２以降の処理を繰り返す。

また、キーワードwが空であれば、そのままステップＳ８２以降の処理を繰り返す。

【０１２４】

そして、ステップＳ８２においてＳ４が空になると、次に、Ｓ６が空かどうかを判定し（ステップＳ８７）、それが空でなければ、ステップＳ８１以降の処理を繰り返す。

【０１２５】

そして、ステップＳ８７においてＳ６が空になると、それ以上処理を繰り返しても新たなパスは設定されないと判断し、その時点でまだパスが設定されていないキーワードの集合をＳ７とする（ステップＳ８８）。そして、Ｓ７の各キーワードのパスのフィールドに、トップの直下のカテゴリ“その他”を設定し、処理を終了する。こうして、キーワードテーブル６２のすべてのキーワードにパスが設定される。

【０１２６】

図１６のようなディレクトリファイル５６が生成されると、文書の分類登録が完了する。利用者は、ディレクトリアクセス部４４および検索部４５を介して、ディレクトリファイル５６の必要な情報を取得することができる。

【０１２７】

ディレクトリアクセス部４４への利用者からの入力としては、表示されたリンクのクリックと検索要求の２種類が考えられる。リンクがクリックされた場合は、ディレクトリアクセス部４４は、対応するファイル１０２、１０３の内容を表示装置１４に表示する。利用者は、図１６のハイパーテキストによる索引を、次のように利用することができる。

【０１２８】

１．パス：ハイパーテキスト全体の中における現在位置が把握できる。迷子にならないための工夫である。

２．上位関連語：キーワードに関連する上位または広い概念のカテゴリとして、検索結果を広げるために用いられる。

【０１２９】

３．サブカテゴリ：キーワードの下位のカテゴリとして、検索結果を絞り込むために用いられる。

４．連想語：キーワードとの関連性は低いが、文書を通じて繋がっているカテゴリとして、ハイパーテキストのブラウジング、ジャンプ等に自由に用いられる。

【０１３０】

また、検索要求に関しては、ディレクトリトップファイル１０１の画面上の入力窓１０４からのキーワード検索と、ディレクトリ中間ファイル１０３の画面上の入力窓１１２からの文書内容検索の２種類がある。ディレクトリアクセス部４４は、これらの検索を検索部４５に指示し、その結果を受け取って表示装置１４に表示する。

【０１３１】

キーワード検索の場合は、検索部４５は、キーワードテーブル６２を検索し、検索要求を満たすキーワードをリストアップする。そして、ディレクトリアクセス部４４は、各キーワードのディレクトリ中間ファイル１０３へのリンクを、ディレクトリトップファイル１０１に付加する。文書内容検索の場合は、検索部４５は、文書データ２１の文書のうち、検索要求を満たす文書をリストアップし、ディレクトリアクセス部４４は、それらのタイトルおよび内容のリストを表示する。

【０１３２】

例えば、利用者が“自動車のワックス”についての情報を探したいと思った場合、まず、“自動車”でキーワード検索を行って文書を絞り込み、次に、“ワックス”で文書内容を検索することで、検索結果のゴミを減らすことができる。

【０１３３】

単に“ワックス”を含む文書を検索しただけでは、床のワックスやスキーのワックスに関する文書等の不要な文書まで検索結果に含まれてしまう。また、“自動車”のような一般的な語は、往々にして、自動車関係の文書にはそのままの形で出現しないことが多いため、“自動車　ＡＮＤ　ワックス”のような検索式でブール検索を行っても、良い結果が得られない場合も多い。

【０１３４】

以上説明したように、文書整理装置は、文書に付加されたキーワードの統計情報、文字列としての包含関係、および辞書等を用いて人手で与えた関係を統合して、ディレクトリを自動的に構築する。統計情報だけでは分類の精度は低く、既存の分類をベースにすると汎用的だが文書の特徴をうまく表現できない。本実施形態では、両者を組み合わせることで、汎用性を保ちつつ、分類の精度を向上させている。

【０１３５】

ディレクトリサービスにおいては、階層関係による分類以外に、上位関連語、連想語、５０音・アルファベット順索引といった多様なリンクを最短パスやサブカテゴリとともに提示することで、文書への複数のパスが提供され、利用者のアクセスが支援される。したがって、このサービスでは、分類そのものより、利用者を文書にナビゲートする多くの手段を提供することに主眼が置かれている。

【０１３６】

また、管理者が同意語・不要語リストを明示的に与え、文書整理装置がその情報に従ってキーワードのリンクを追加・削除することで、管理者の意向をディレクトリに反映させることができる。また、管理者がキーワード間の階層関係を明示的に与え、文書整理装置がその情報に従ってディレクトリを生成することで、ハイパーテキストのリンク関係を調整することができる。

【０１３７】

また、文書整理装置は、今回入力されたキーワードを前回のキーワードと比較し、新規に登録されたキーワードを強調表示するため、管理者の手間をかけずに、利用者が新しい話題を把握するための手掛かりが提供される。

【０１３８】

また、文書整理装置は、文書のキーワードをそのままディレクトリのカテゴリとして用いるため、従来の自動分類とは違って、最初から分類ミスが生じない。さらに、文書本文の全文検索と、ディレクトリの分類とを融合することで、話題を絞りこんで検索することが可能となる。これにより、同音異議語による検索ゴミを減らすことができる。

【０１３９】

また、ディレクトリに含まれるキーワードの検索と文書内容の全文検索を組み合わせることで、利用者が特定の話題における文書から細かい情報を検索することが支援される。

【０１４０】

次に、図２０から図２４までを参照しながら、図２に示した文書整理装置を利用した文書整理システムの実施形態について説明する。今日のネットワーク環境においては、メールやニュースといった身の回りの文書が、コンピュータ上の文書フォルダに溜まっていることが多い。文書整理システムは、そのような文書群に対して本発明を応用し、文書整理を行う。

【０１４１】

図２０は、このような文書整理システムの構成図である。図２０の文書整理システムは、処理装置１２１、二次記憶装置１２２、キーワード抽出装置１２３、および利用者端末１２４を備える。例えば、処理装置１２１は、ＣＰＵとメモリを含み、利用者端末１２４は、入力装置と表示装置を含む。

【０１４２】

二次記憶装置１２２は、文書群のデータを含む文書フォルダ１３１と管理ファイル１３２を格納する。管理ファイル１３２には、同意語および不要語の集合１３３と、キーワードの階層関係を表すデータ１３４と、ディレクトリのトップとなるキーワードの集合１３５が含まれている。

【０１４３】

キーワード抽出装置１２３は、文書フォルダ１３１の各文書の形態素解析を行って、文書を単語に分割する。そして、中頻度の単語をキーワードとして取り出し、処理装置１２１に入力する。低頻度の単語では文書の特徴を表していない恐れがあり、高頻度の単語では他の文書にも多く現れる可能性がある。

【０１４４】

処理装置１２１は、キーワード整形器４１、キーワード関係抽出器４２、ディレクトリファイル生成器４３、検索部４５、ＷＷＷ（world wide web）サーバ１４１を含む。

【０１４５】

キーワード整形器４１、キーワード関係抽出器４２、およびディレクトリファイル生成器４３は、文書フォルダ１３１および管理ファイル１３２のデータと、キーワード抽出装置１２３からのキーワードとを用いて上述したような処理を行い、図１６に示したような形式のディレクトリファイル１４２を生成する。

【０１４６】

また、WWWサーバ１４１は、図２のディレクトリアクセス部４４に対応し、利用者からの指示に応じてディレクトリファイル１４２にアクセスする。利用者は、端末１２４に搭載されたWWWブラウザ１４３を通じて、WWWサーバ１４１に対する指示を入力し、ディレクトリにアクセスする。

【０１４７】

図２１は、端末１２４上に表示されるディレクトリのトップ画面を示している。トップ画面の項目“コンピュータ”、“ソフトウェア”等は、管理者がトップキーワード１３５により与えたカテゴリに相当する。

【０１４８】

図２２は、利用者がトップ画面または他の画面上のキーワードをクリックすることにより表示されるディレクトリの中間画面を示している。ここでは、“ホームページ”というキーワードのページが表示されている。

【０１４９】

右上のパス１５１は、トップ画面から“ホームページ”の中間画面に至るパスを表し、この中間画面はトップ画面の下位のキーワード“WWW”の下位に位置することが分かる。また、関連語１５２は、“ページ”、“インターネット”等が“ホームページ”の上位語であることを表し、サブトピック（サブカテゴリ）１５３は、“WWWページ”、“接続事業”、“ＨＴＭＬ”等の１２個のキーワードが“ホームページ”の下位語であることを表している。

【０１５０】

また、文書リスト１５４は、“ホームページ”に関連する３９個の文書のそれぞれについて、タイトル、本文（一次情報）へのリンク、更新日、および連想語のリンクを示している。例えば、最初の文書タイトル“国立天文台のWWWペー

ジ”の下の括弧内に記述された“日本”、“電子メール”、および“代理投稿”が、“ホームページ”の連想語である。

【０１５１】

図２３は、ディレクトリの５０音・アルファベット順索引のトップ画面を示しており、図２４は、その下の中間画面を示している。図２４では、“れ”で始まるキーワードがリストアップされている。このような文書整理システムによれば、任意の文書を整理して格納することができる。

【０１５２】

また、本実施形態の文書整理装置は、他にも次のようなシステムに応用することができる。

（１）情報共有ツールのビュー

本出願人による先願である「文書共有整理システム、共有文書管理装置および文書アクセス装置」（特願平８－２８１９４０）では、ネットワークを通じてグループで文書情報を共有し、特定の文書のリストとしてビューを表示することができる。このビューの１つとして、文書整理装置が作成するディレクトリを表示することが考えられる。

（２）ネットワークニュース検索システム

本出願人による先願である「関連文書表示装置」（特願平１０－８２２７０）では、ネットワークニュースの検索システムが開示されている。ネットワークニュースにおけるニュースグループの整理に文書整理装置を応用することで、利用者のアクセス支援ができる。

【０１５３】

なお、以上説明した実施形態においては、管理者が管理ファイルを作成しているが、利用者自身が管理者の役割を兼ねることもあり得る。また、本発明は、文書のみならず、キーワードを付加されたあらゆる情報の分類・整理に応用することが可能である。例えば、画像や音声のファイルに適当なキーワードを付加しておき、それらの間の関係からディレクトリファイルを作成することができる。

【０１５４】

ところで、図２の文書整理装置は、図２５に示すような情報処理装置（コンピ

ュータ）を用いて構成することができる。図２５の情報処理装置は、ＣＰＵ１６１、メモリ１６２、入力装置１６３、出力装置１６４、外部記憶装置１６５、媒体駆動装置１６６、およびネットワーク接続装置１６７を備え、それらはバス１６８により互いに接続されている。

【０１５５】

メモリ１６２は、例えば、ＲＯＭ（read only memory）、ＲＡＭ（random access memory）等を含み、処理に用いられるプログラムとデータを格納する。ＣＰＵ１６１は、メモリ１６２を利用してプログラムを実行することにより、必要な処理を行う。

【０１５６】

図２のキーワード整形器４１、キーワード関係抽出器４２、ディレクトリファイル生成器４３、ディレクトリアクセス部４４、および検索部４５は、それぞれ、メモリ１６２の特定のプログラムコードセグメントにプログラムとして格納される。

【０１５７】

入力装置１６３は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置１６４は、例えば、ディスプレイやプリンタ等であり、利用者への問い合わせ、処理結果等の出力に用いられる。

【０１５８】

外部記憶装置１６５は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク（magneto-optical disk）装置等であり、図２の二次記憶装置１２として用いられる。この外部記憶装置１６５に、上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ１６２にロードして使用することもできる。

【０１５９】

媒体駆動装置１６６は、可搬記録媒体１６９を駆動し、その記録内容にアクセスする。可搬記録媒体１６９としては、メモリカード、フロッピーディスク、ＣＤ－ＲＯＭ（compact disk read only memory）、光ディスク、光磁気ディスク

等、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体１６９に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ１６２にロードして使用することもできる。

【０１６０】

ネットワーク接続装置１６７は、ＬＡＮ（local area network）等の任意のネットワーク（回線）を介して外部の装置と通信し、通信に伴うデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ１６２にロードして使用することもできる。

【０１６１】

図２６は、図２５の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体１６９や外部のデータベース１７０に保存されたプログラムとデータは、メモリ１６２にロードされる。そして、ＣＰＵ１６１は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【０１６２】

【発明の効果】

本発明によれば、情報処理装置に蓄えられた大量の文書群を、外部から与えられたカテゴリと自動的に抽出された文書群の特徴を併用して、高い精度で自動的に分類することができる。また、分類結果に従って、文書への多様なリンクを有するディレクトリが自動的に生成され、利用者によるアクセスが支援される。

【図面の簡単な説明】

【図１】

本発明の文書整理装置の原理図である。

【図２】

文書整理装置の構成図である。

【図３】

管理ファイルのデータ構造を示す図である。

【図４】

文書単語対テーブルとキーワードテーブルを示す図である。

【図５】

文書メタ情報のデータ構造を示す図である。

【図６】

キーワード整形器の構成図である。

【図７】

キーワード整形器の処理のフローチャートである。

【図８】

キーワード関係抽出器の構成図である。

【図９】

キーワード関係抽出器の処理のフローチャートである。

【図１０】

相関ルール抽出／評価処理のフローチャートである。

【図１１】

ルール分割を示す図である。

【図１２】

第１の文書集合の関係を示す図である。

【図１３】

第２の文書集合の関係を示す図である。

【図１４】

第３の文書集合の関係を示す図である。

【図１５】

第４の文書集合の関係を示す図である。

【図１６】

ディレクトリファイルを示す図である。

【図１７】

ディレクトリファイル生成器の処理のフローチャートである。

【図１８】

パス生成処理のフローチャート（その１）である。

【図１９】

パス生成処理のフローチャート（その２）である。

【図２０】

文書整理システムの構成図である。

【図２１】

文書ディレクトリのトップ画面を示す図である。

【図２２】

文書ディレクトリの中間画面を示す図である。

【図２３】

文書５０音・アルファベット順索引のトップ画面を示す図である。

【図２４】

文書５０音・アルファベット順索引の中間画面を示す図である。

【図２５】

情報処理装置の構成図である。

【図２６】

記録媒体を示す図である。

【符号の説明】

１　関係抽出手段

２　生成手段

３　出力手段

４、３２、５３、１３４　階層関係

５、５５　連想関係

１１、１２１　処理装置

１２、１２２　二次記憶装置

１３　入力装置

１４　表示装置

２１　文書データ

２２、１３２　管理ファイル

３１、１３３　同意語／不要語

３３、１３５　トップキーワード

４１　キーワード整形器

４２　キーワード関係抽出器

４３　ディレクトリファイル生成器

４４　ディレクトリアクセス部

４５　検索部

５１　文書単語対

５２　文書メタ情報

５４　同値関係

５６、１４２　ディレクトリファイル

６１　文書単語対テーブル

６２　キーワードテーブル

７１　キーワード統一器

７２　単語読み付加器

７３　旧キーワード

８１　部分文字列切り出し器

８２　相関ルール抽出器

８３　ルール評価器

８４　マージャ

８５　部分語関係

８６　相関ルール

９１、９２、９３、９４　領域

１０１　ディレクトリトップファイル

１０２　５０音・アルファベット順索引中間ファイル

１０３　ディレクトリ中間ファイル

１０４、１１２　検索窓

１０５　キーワード

１０６　５０音索引

１０７　ヘッダ

１０８、１５１　パス

１０９、１５２　上位関連語

１１０、１５３　サブカテゴリ

１１１、１５４　文書リスト

１２３　キーワード抽出装置

１２４　利用者端末

１３１　文書フォルダ

１４１　WWWサーバ

１４３　WWWブラウザ

１６１　ＣＰＵ

１６２　メモリ

１６３　入力装置

１６４　出力装置

１６５　外部記憶装置

１６６　媒体駆動装置

１６７　ネットワーク接続装置

１６８　バス

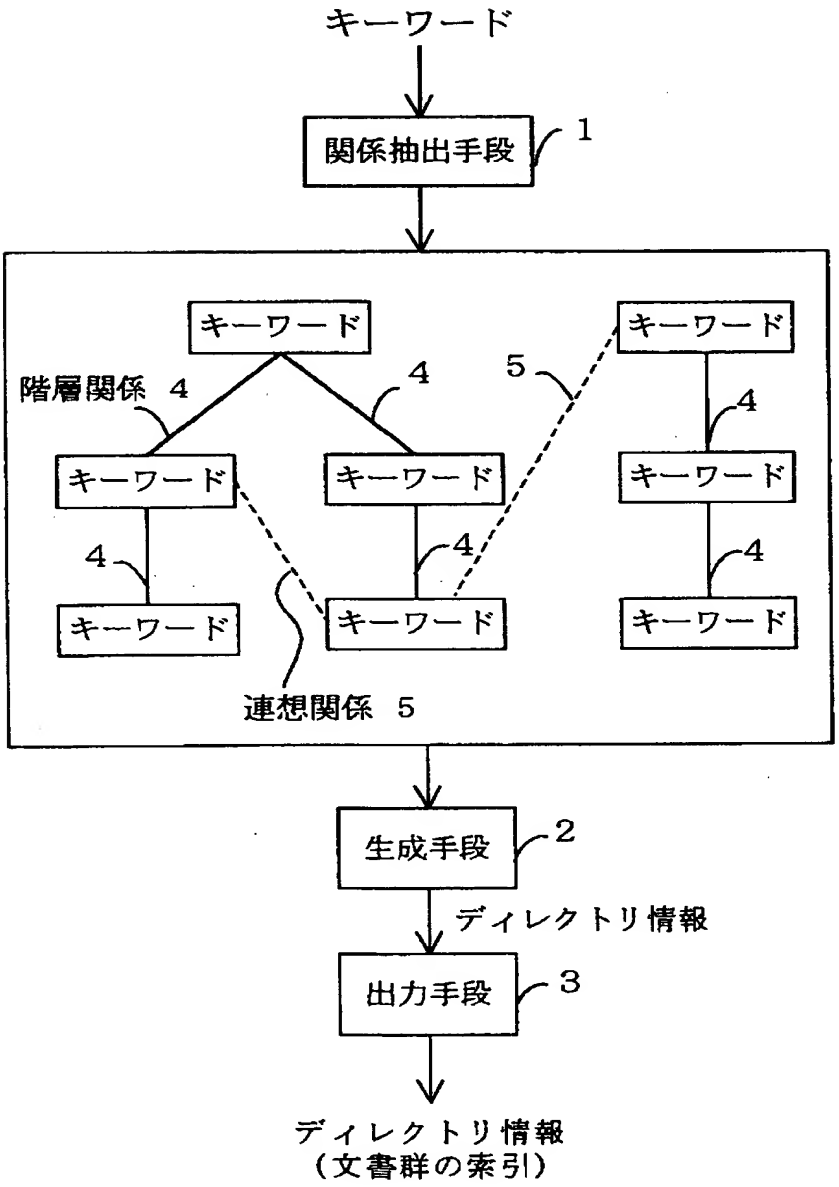
１６９　可搬記録媒体

１７０　データベース

【書類名】　　図面

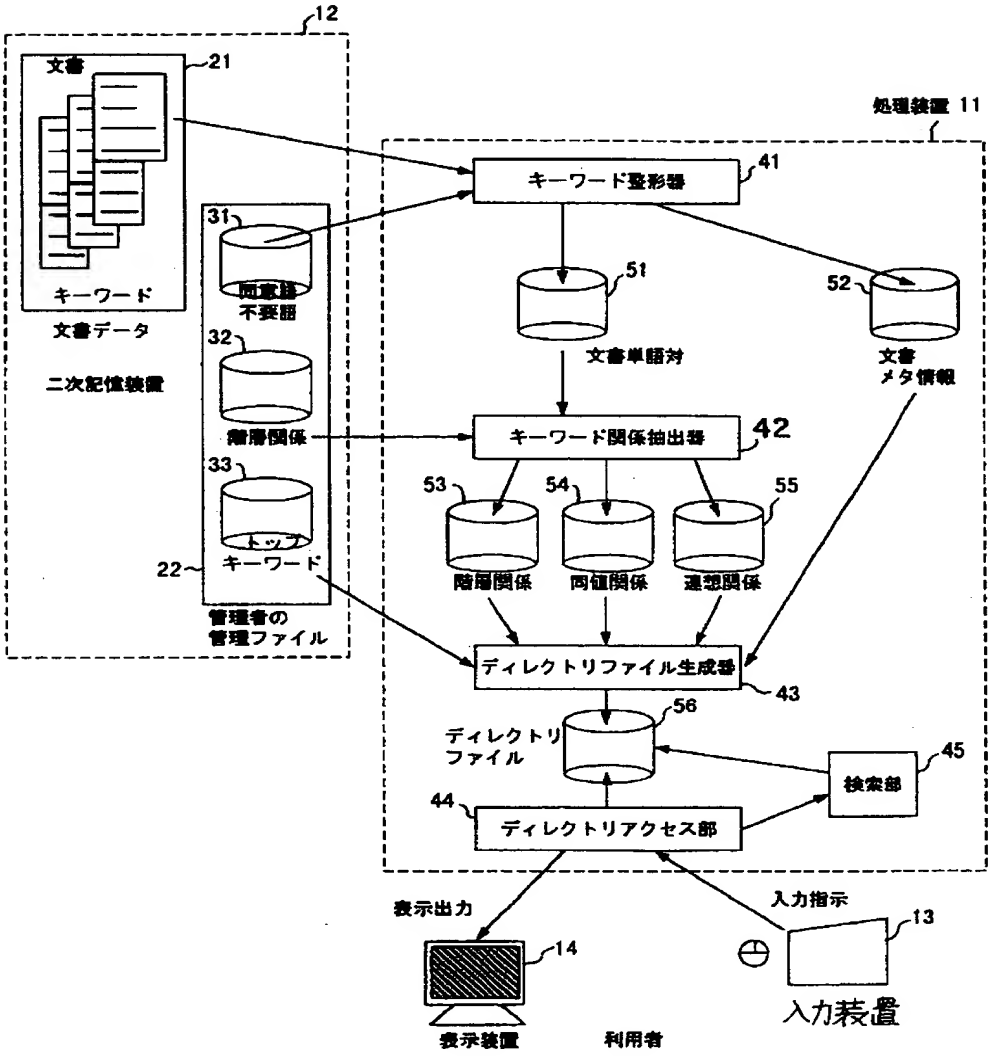
【図１】

本　発　明　の　原　理　図


キーワード
関係抽出手段
1
キーワード
キーワード
階層関係 4
4
5
4
キーワード
キーワード
キーワード
4
4
4
キーワード
キーワード
キーワード
連想関係 5
生成手段
2
ディレクトリ情報
出力手段
3
ディレクトリ情報
（文書群の索引）

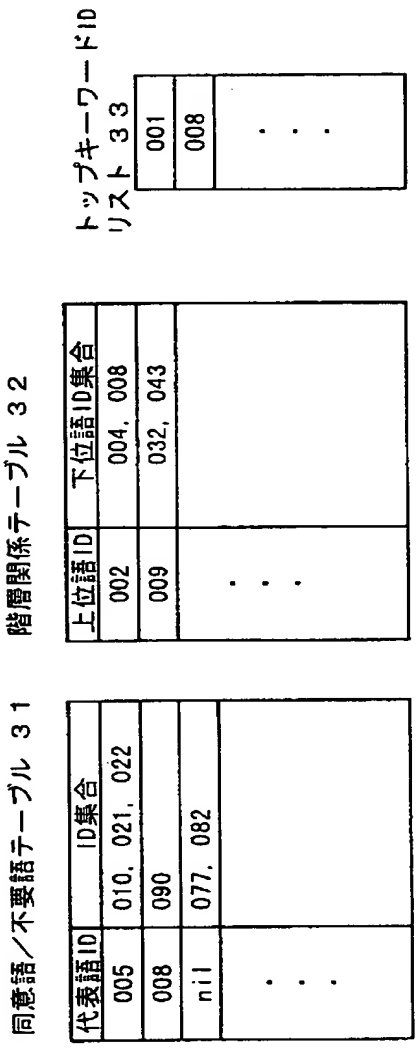
【図２】

文書整理装置の構成図



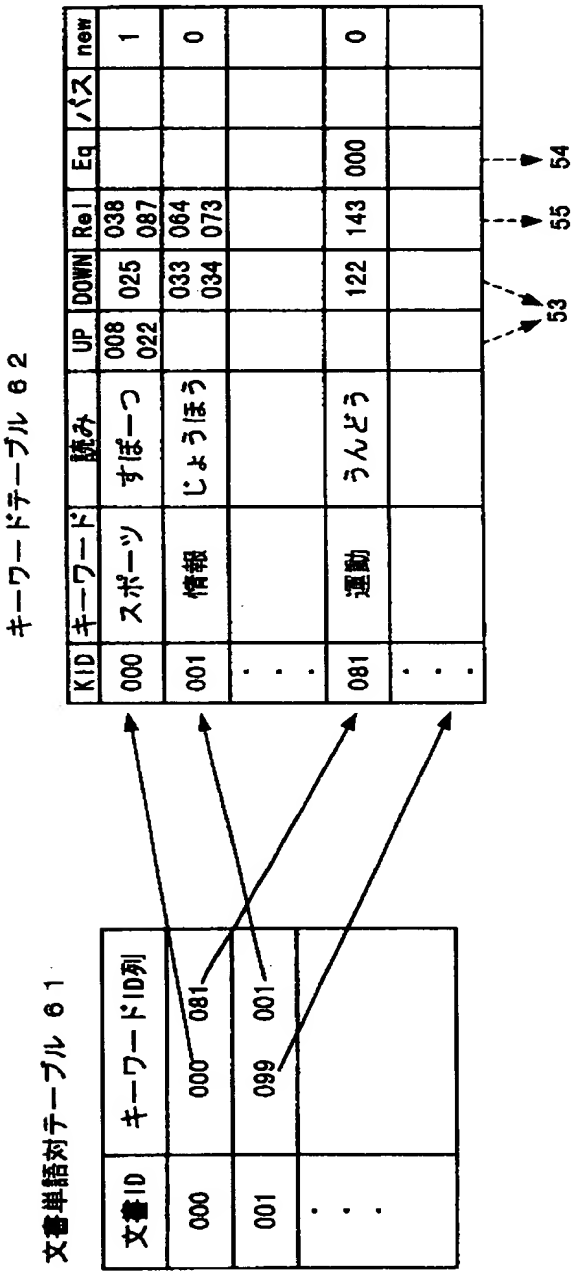
【図３】

管理ファイルのデータ構造を示す図



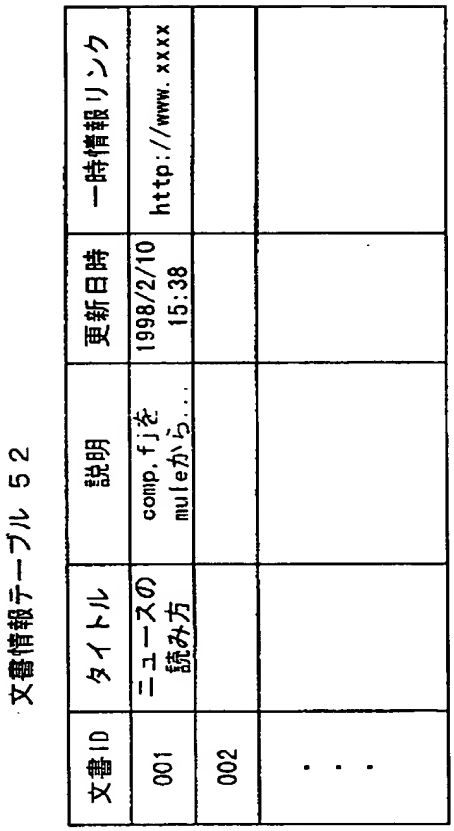
【図４】

文書単語対テーブルとキーワードテーブルを示す図



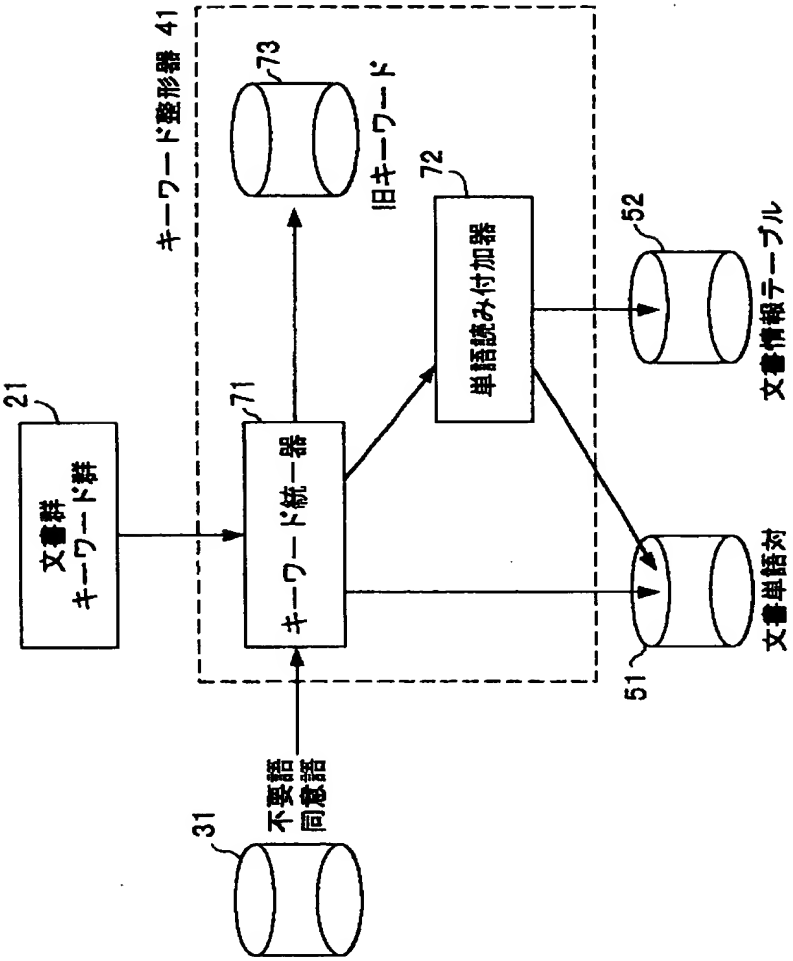
【図５】

文書メタ情報のデータ構造を示す図



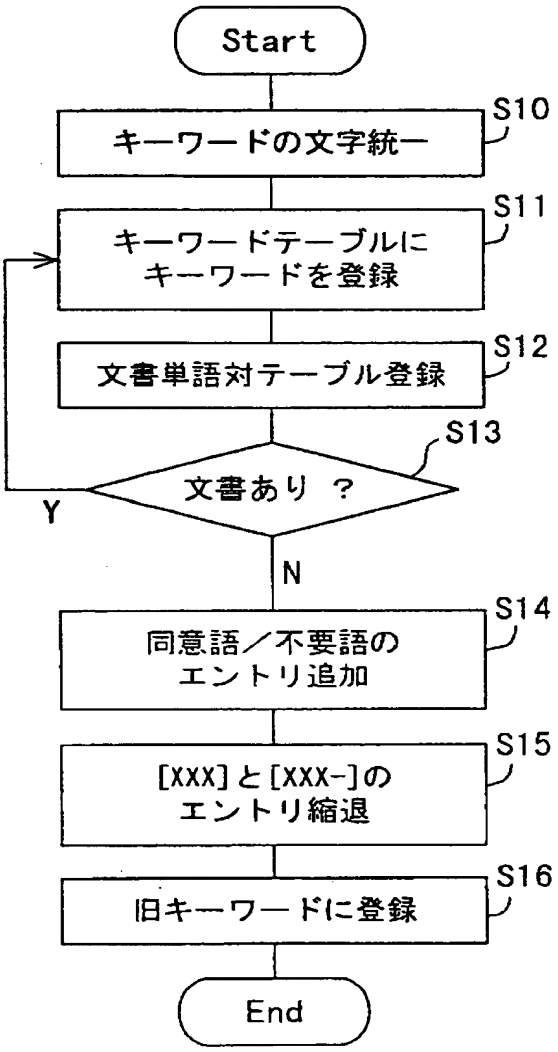
【図６】

キーワード整形器の構成図



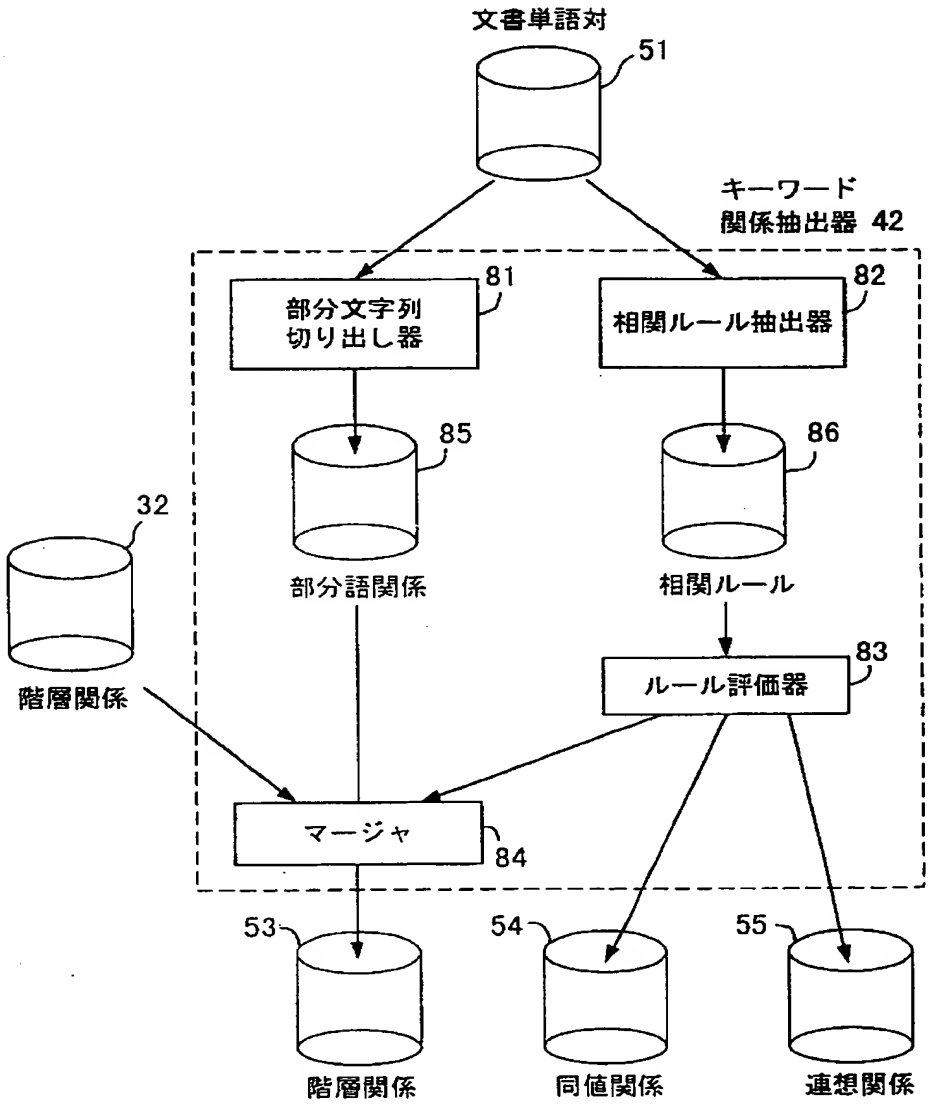
【図７】

キーワード整形器の処理のフローチャート



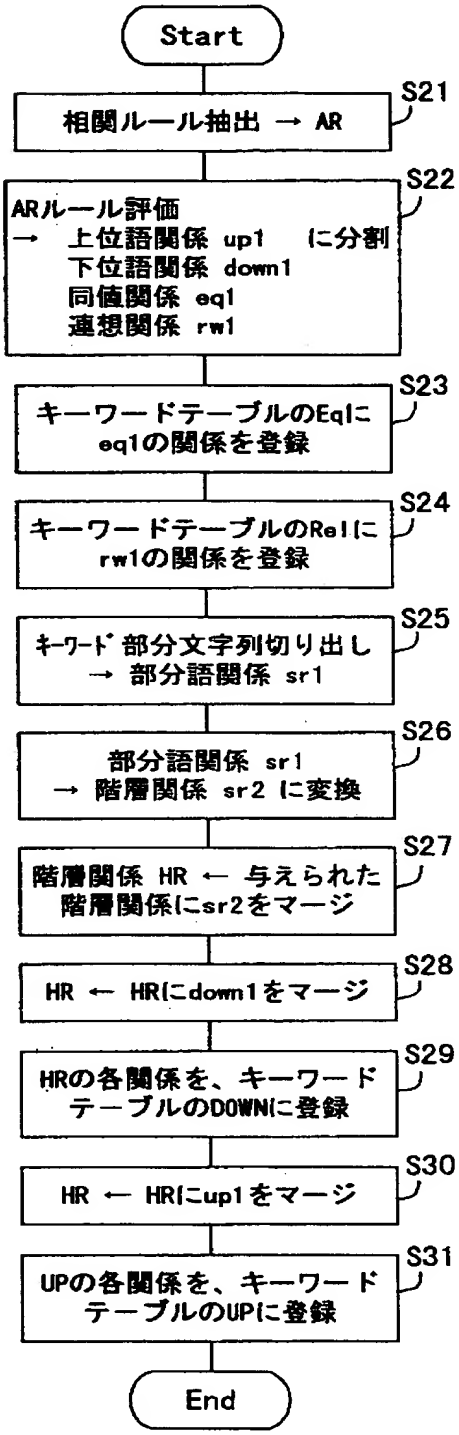
【図８】

キーワード関係抽出器の構成図



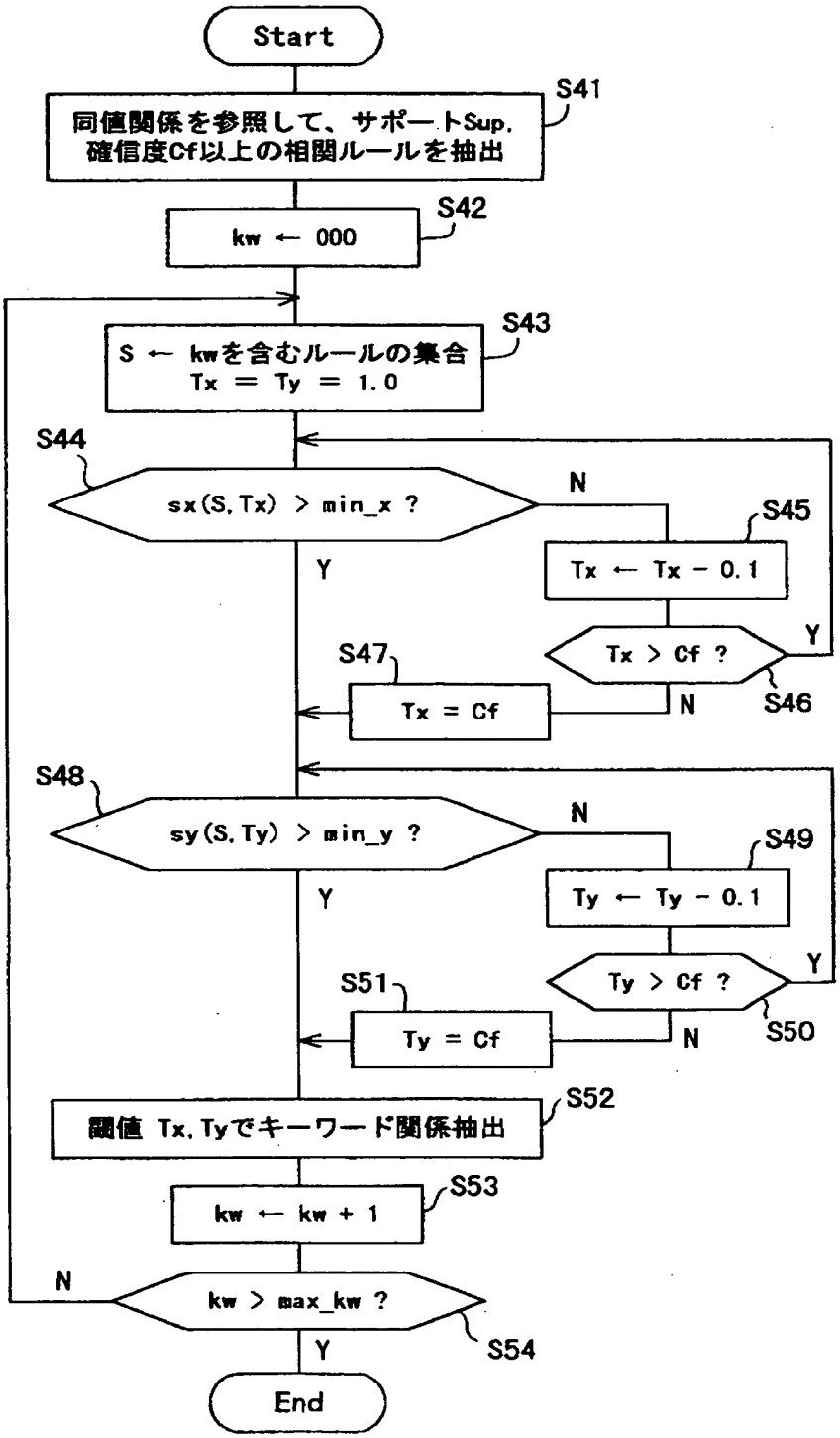
【図９】

キーワード関係抽出器の処理のフローチャート



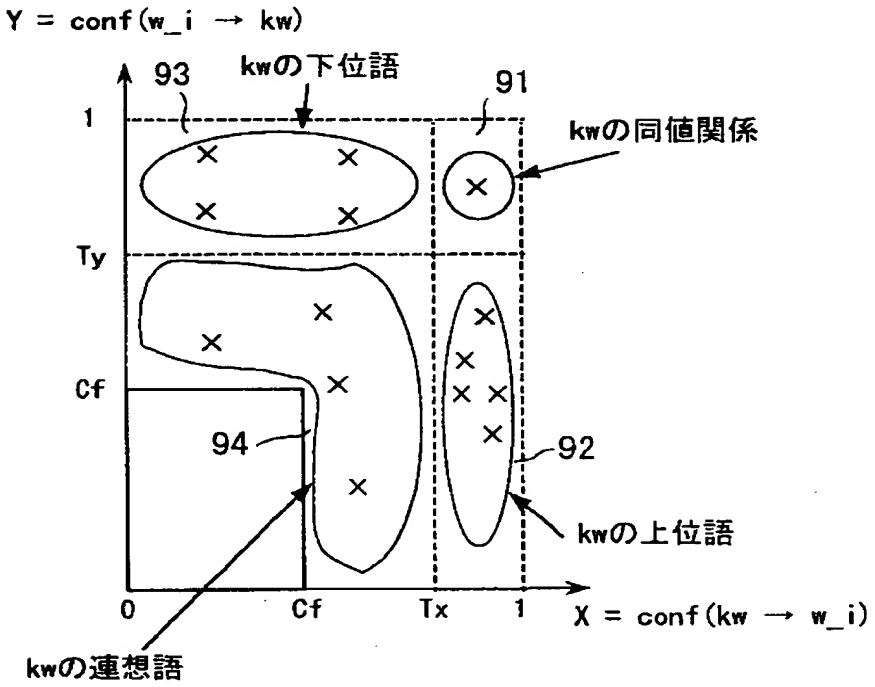
【図１０】

相関ルール抽出／評価処理のフローチャート



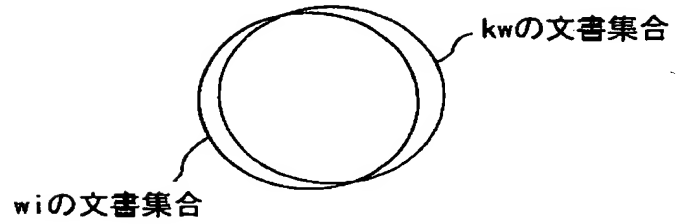
【図１１】

ルール分割を示す図



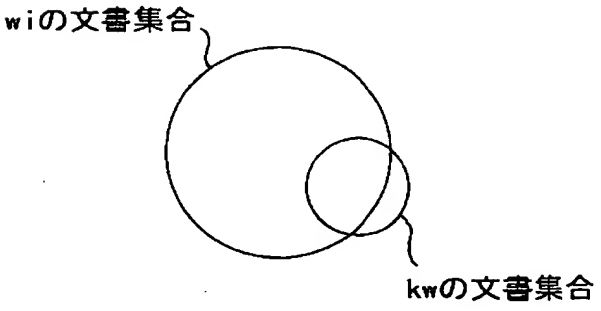
【図１２】

第１の文書集合の関係を示す図



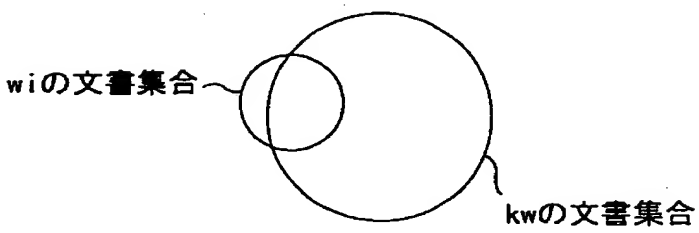
【図１３】

第２の文書集合の関係を示す図



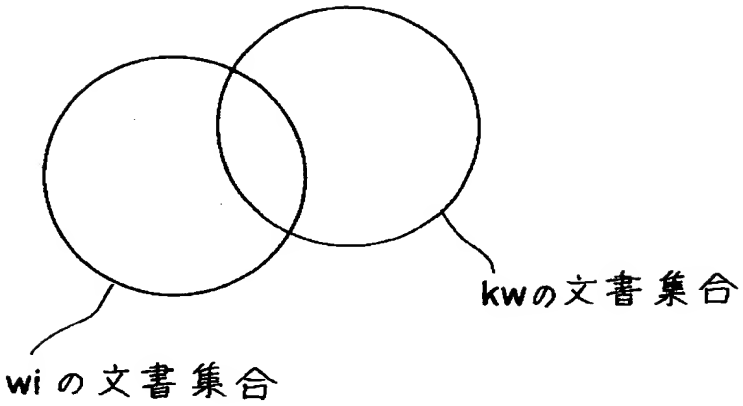
【図１４】

第３の文書集合の関係を示す図



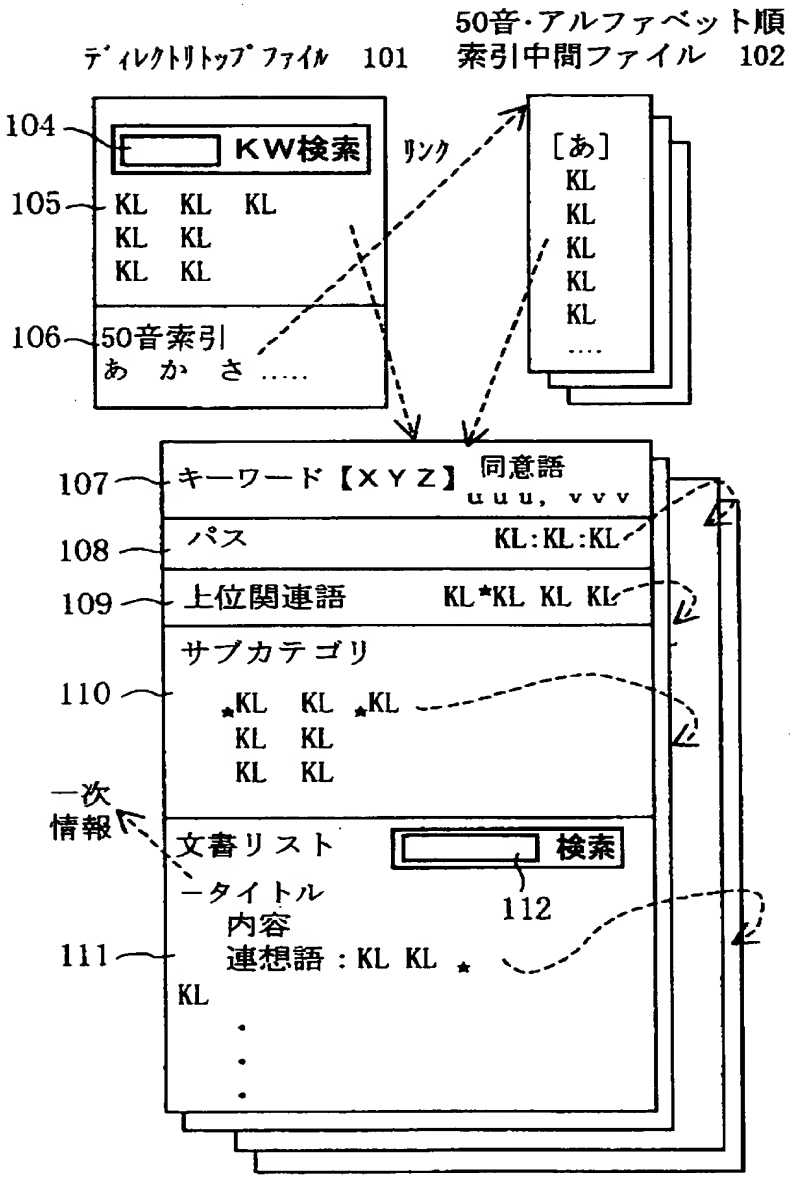
【図１５】

第4の文書集合の関係を示す図



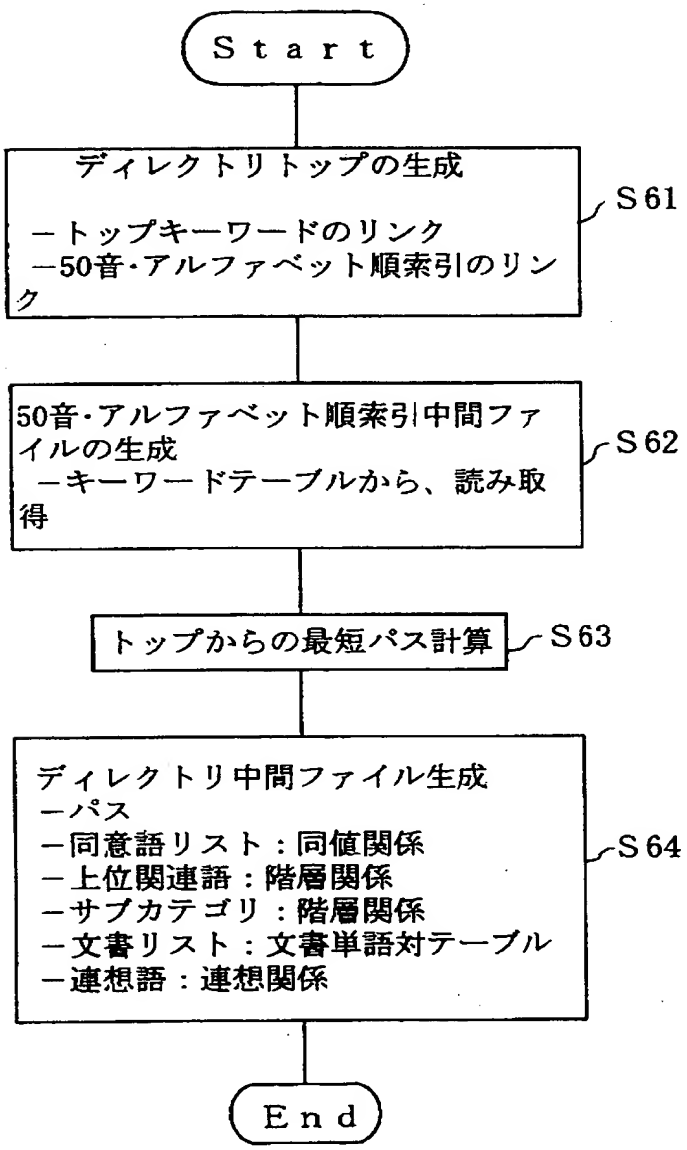
【図１６】

ディレクトリファイルを示す図



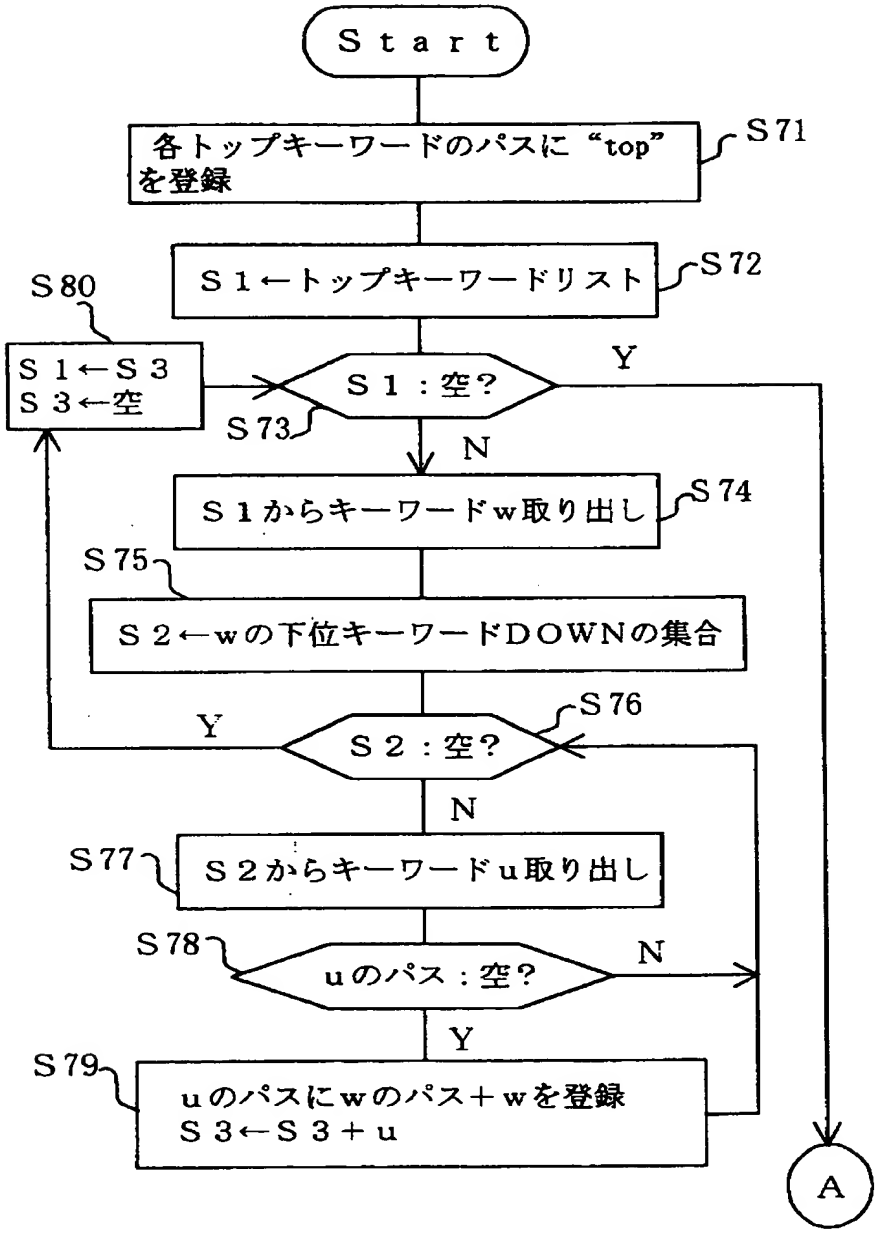
【図１７】

ディレクトリファイル生成器の
処理のフローチャート



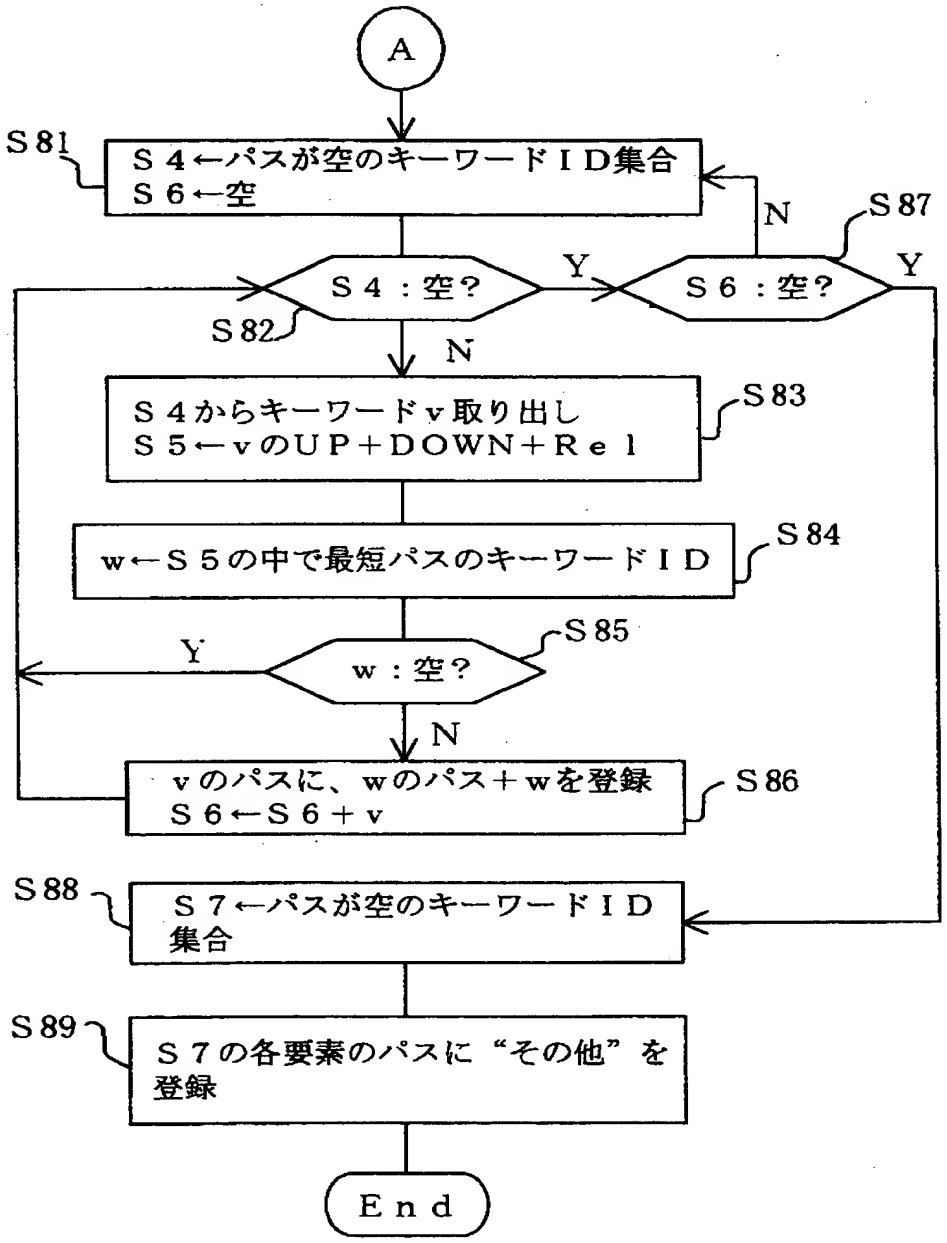
【図１８】

パス生成処理のフローチャート（その１）



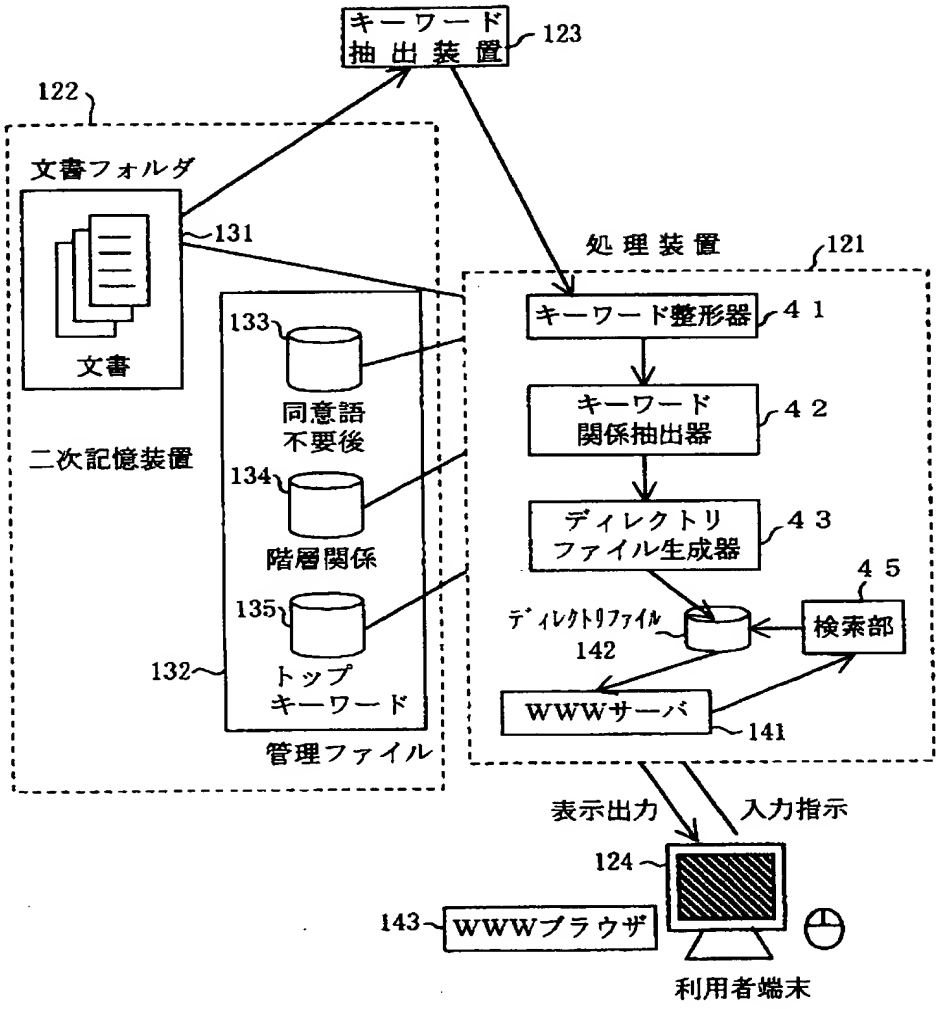
【図１９】

パス生成処理のフローチャート（その２）



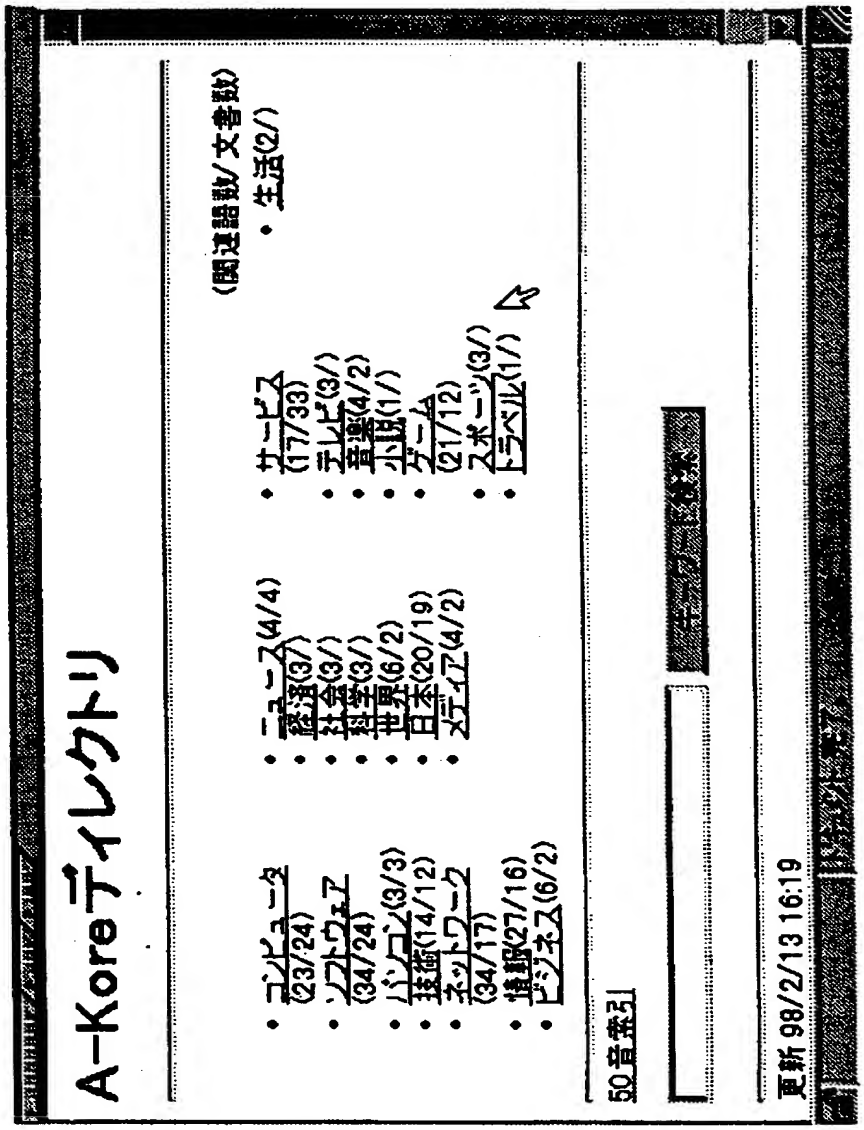
【図２０】

文書整理システムの構成図



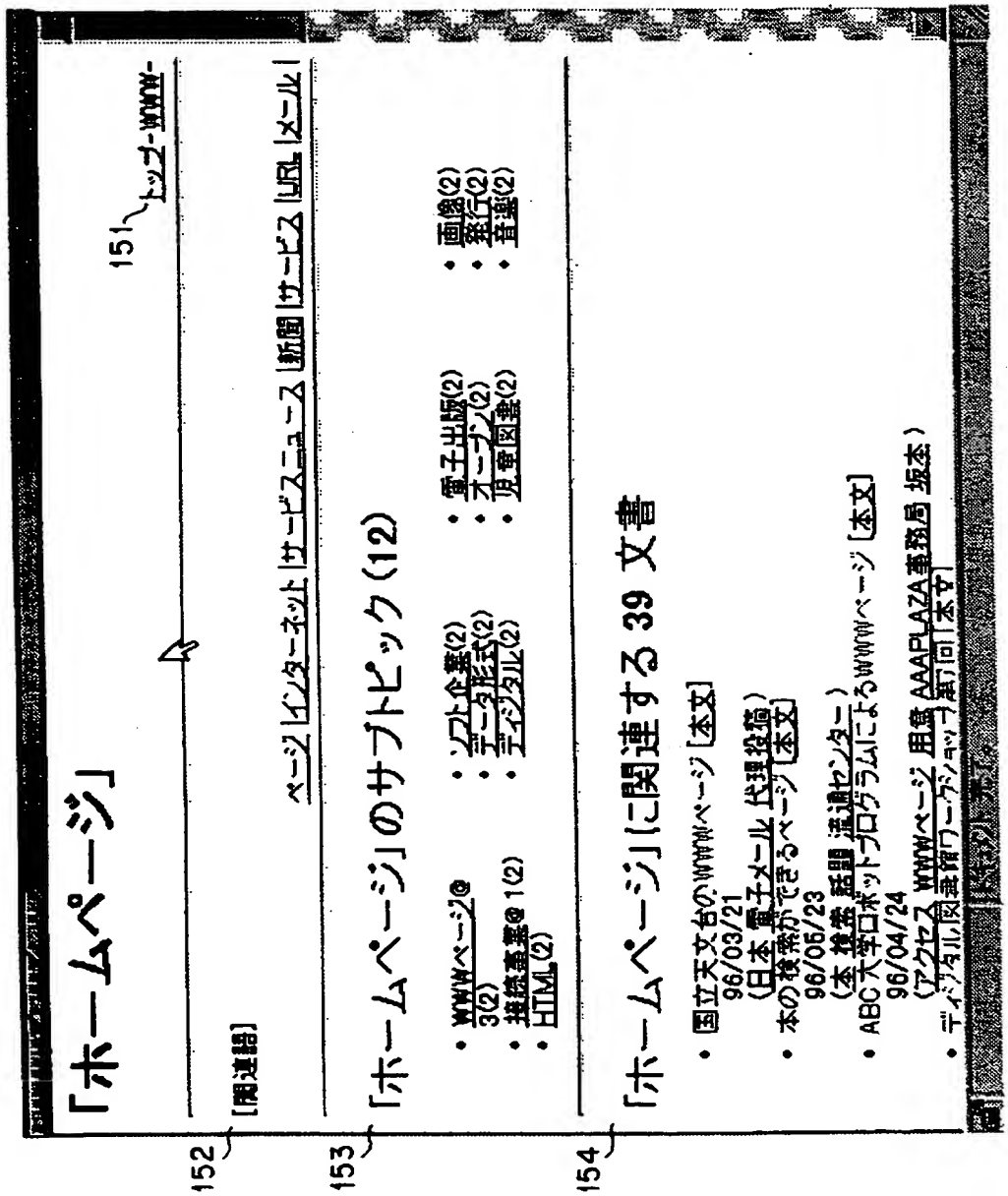
【図２１】

文書ディレクトリのトップ画面を示す図



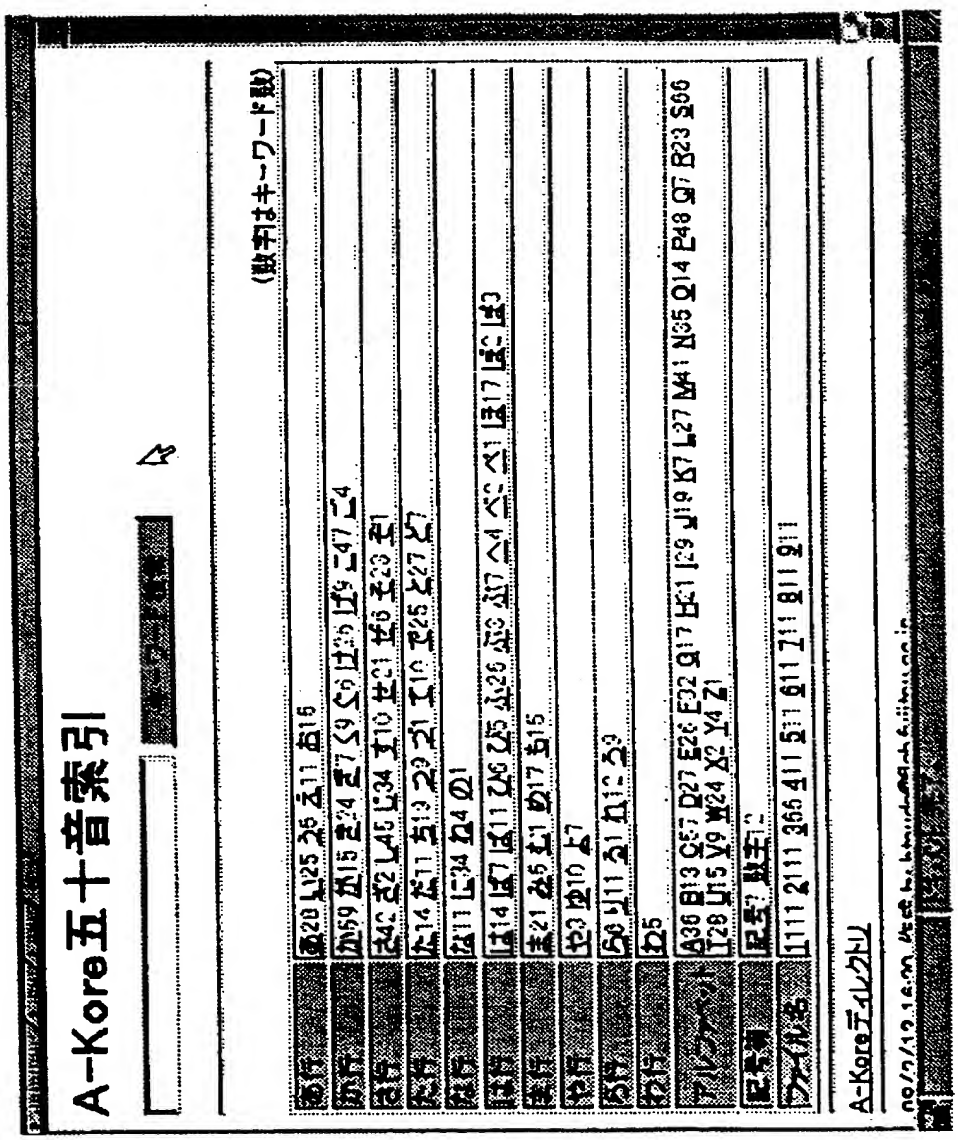
【図２２】

文書ディレクトリの中間画面を示す図



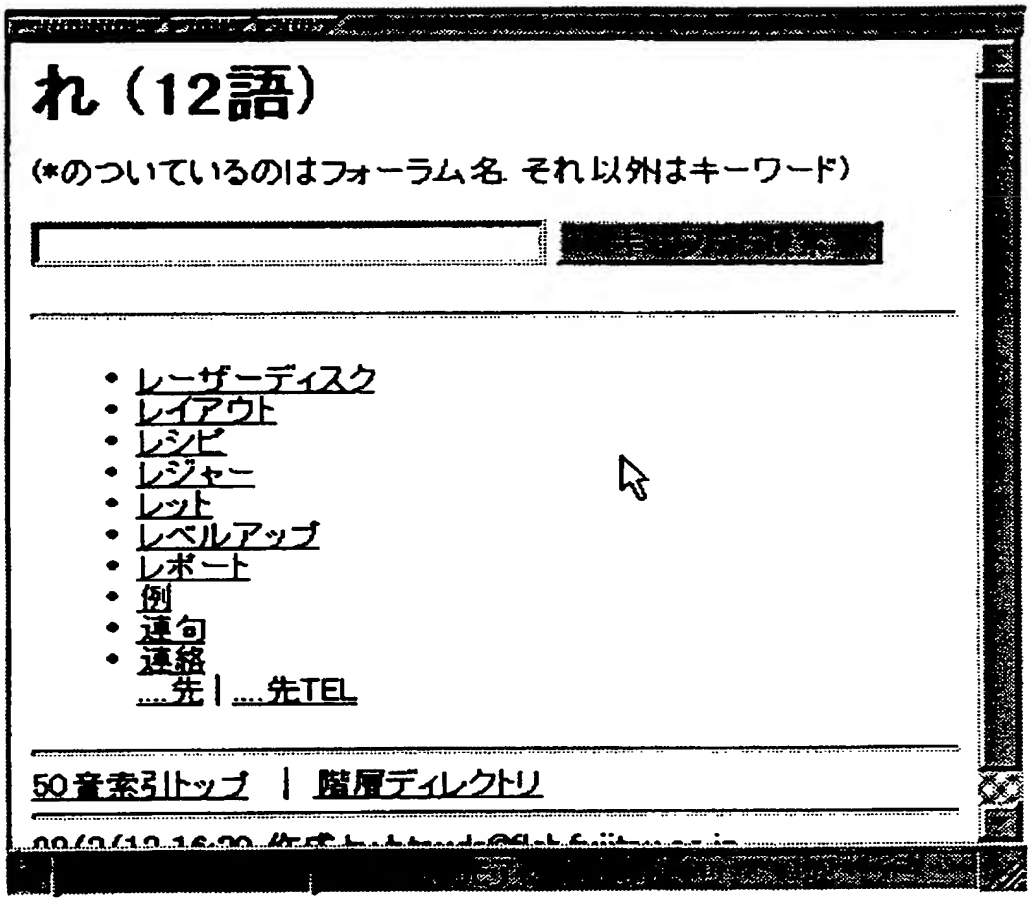
【図２３】

文書50音・アルファベット順索引のトップ画面を示す図



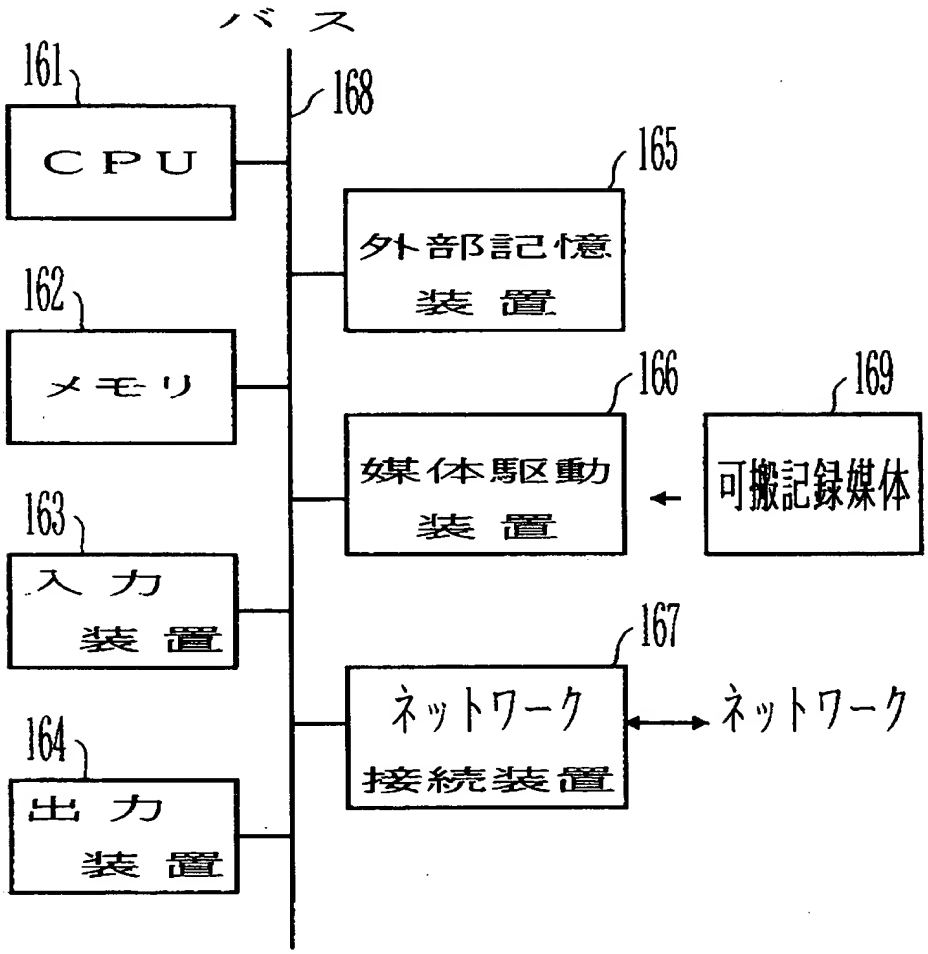
【図２４】

文書50音・アルファベット順索引の中間画面を示す図



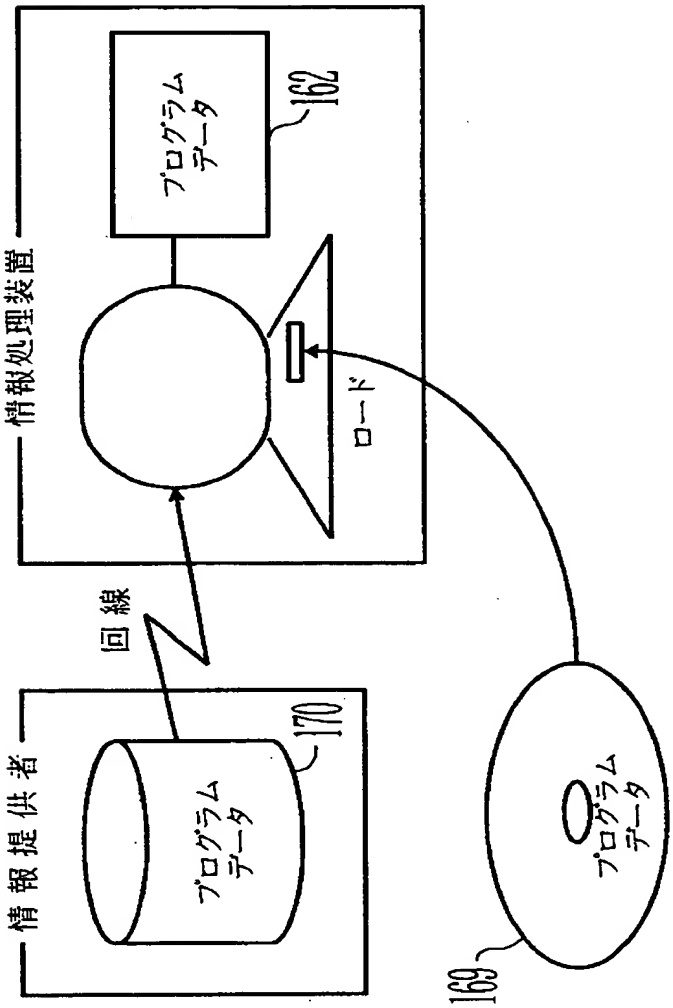
【図２５】

情報処理装置の構成図



【図２６】

記録媒体を示す図



【書類名】　要約書

【要約】

【課題】　情報処理装置に蓄えられた大量の文書群を、その特徴に従って高い精度で自動的に分類することが課題である。

【解決手段】　キーワード整形器４１は、同意語・不要語３１を参照して、文書データ２１から文書単語対５１と文書メタ情報５２を生成する。キーワード関係抽出器４２は、階層関係３２を参照して、階層関係５３、同値関係５４、連想関係５５を抽出し、ディレクトリファイル生成器４３は、トップキーワード３３、階層関係５３、同値関係５４、連想関係５５、および文書メタ情報５２からディレクトリファイル５６を生成する。利用者は、ディレクトリアクセス部４４を介してディレクトリファイル５６にアクセスする。

【選択図】　図２

【書類名】　職権訂正データ
【訂正書類】　特許願

＜認定情報・付加情報＞
【特許出願人】
【識別番号】　000005223
【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号
【氏名又は名称】　富士通株式会社
【代理人】　申請人
【識別番号】　100074099
【住所又は居所】　東京都千代田区二番町８番地２０　二番町ビル３Ｆ　大菅内外国特許事務所
【氏名又は名称】　大菅　義之
【選任した代理人】
【識別番号】　100067987
【住所又は居所】　神奈川県横浜市港北区太尾町１４１８－３０５　（大倉山二番館）　久木元特許事務所
【氏名又は名称】　久木元　彰

出　願　人　履　歴　情　報

識別番号　　〔０００００５２２３〕

１．変更年月日　　１９９６年　３月２６日
［変更理由］　住所変更
住　所　神奈川県川崎市中原区上小田中４丁目１番１号
氏　名　富士通株式会社